ЛИНГВИСТИКА

A. Arkhipov, C. L. Däbritz

HAMBURG CORPORA FOR INDIGENOUS NORTHERN EURASIAN LANGUAGES[1]

The long-term INEL project (2016–2033), carried out at the University of Hamburg, aims to develop digital linguistic corpora and supporting infrastructure for a number of selected languages of Northern Eurasia. At present, corpora of Selkup, Kamas and Dolgan are being created. The project builds upon existing materials from various archive sources, including the Selkup archive of Angelina I. Kuzmina preserved at the University of Hamburg, Kamas audio recordings from the archives in Tartu and Helsinki, and Dolgan recordings provided by the House of the Cultures of Taimyr Peninsula. All the texts in the corpora are provided with a phonological transcription, morphological interlinear glossing, free translations; selected subsets also bear additional annotations for semantic and syntactic features, information status of referents, borrowings and code-switching. The corpora are intended for typologically aware grammatical research but may also be of interest for a wider audience. A number of satellite information resources are also being developed, contributing towards a more efficient research infrastructure.

Key words: *INEL project, corpora, Selkup, Kamas, Dolgan, archive materials, grammar, language contacts, interlinear glossing, annotation.*

1. The INEL project[2]

In 2016, a long-term research project started in Hamburg, Germany, which is to produce a number of corpora of endangered languages and varieties of Northern Eurasia, mainly of Western Siberia.

The INEL project is scheduled for 18 years (2016–2033), allowing 3 years of work for each language or variety. Three subprojects are currently underway, presented in more detail below: Selkup, Kamas (both Samoyedic) and Dolgan (Turkic). The main goal of the project consists in developing digital text corpora for selected lesser described and underresourced languages. The selection is conditioned by the ability to contrast varieties of the same language in different contact situations. For instance, while most Selkup varieties had contacts with Khanty, Evenki and Ket, Northern Selkup in particular had relations with the Nenets, and Southern Selkup (as well as Kamas) had recent contacts with neighboring Turkic languages.

Another factor of selection was the availability of a substantial amount of primary data which would make a significant contribution to the already existing resources on these varieties. Our intention is to mobilize, wherever possible, previously collected data, both in written and spoken form, which might be available from institutional or personal archives. INEL is thus to be seen not as a language documentation project, but as one building upon the existing data of different origin to

[1] This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

[2] "Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages" (INEL). Principal Investigator: Prof. Dr. Beáta Wagner-Nagy. Application by Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, MA Hanna Hedeland, MA Timm Lehmberg.
Current project members: Alexandre Arkhipov, Maria Brykina, Chris Lasse Däbritz, Valentin Gusev and Svetlana Orlova (linguistic team); Timm Lehmberg, Anne Ferger and Daniel Jettka (development team). Former project members: Josefina Budzisch, Tiina Klooster, Niko Partanen and Hannah Wegener. Project website: URL: https://inel.corpora.uni-hamburg.de/

transform them into state-of-the-art digital annotated corpora, ready to use in typologically aware grammar-oriented research and grammatical description.

2. Data processing

The core steps of linguistic analysis for all languages are morphological glossing and (partial) annotation of several further aspects of grammatical structure, borrowings and code-switching. These core steps are preceded by technical preparation of the source data, according to their origin and nature. Steps taken for transcribing and translating the texts (prior to glossing) may also differ depending on the sources. For instance, Selkup texts from A. Kuzmina's manuscript fieldnotes come in an already transcribed and partly translated form, but they still need conversion to the transcription system used in the project, and the fragmentary Russian translation needs to be completed and rewritten. Other translation languages are added in parallel with or after the glossing. Several revision cycles are often needed to harmonize the final translations with the text analysis and across translation languages (English, German and Russian).

In the case of sound recordings, **transcription** of sound files and primary **translation** can be done (i) by a linguist alone, (ii) by a linguist assisted by a native speaker (in fieldwork sessions), or, when an educated and computer-trained native speaker is available, also (iii) remotely by a native speaker alone. The latter way is not often possible and is the most demanding in terms of personal and technical requirements, initial preparation, workflow organization, and post-processing. However, for larger volumes of data it is much more efficient in terms of output per researcher's working hour. At present this approach could be partially implemented for Dolgan as well as Central and Southern Selkup data.

Morphological glossing follows the principles of the Leipzig Glossing Rules (), de-facto standard in typology and language documentation. Along with the glossing labels, grammatical categories for each stem and affix and part-of-speech labels for each word are provided.

Several **further tiers of annotations** are added for a subset of the glossed texts:

1. Syntactic functions (SyF): for predicates and arguments.
2. Semantic roles (SeR): for principal arguments and some adjuncts.
3. Information status (IST): status of referential expressions (e.g. noun phrases, pronouns, null arguments) as given/accessible/new.
4. Borrowings (BOR, BOR-Phon, BOR-Morph): source language, lexical category, phonological and morphological adaptations of borrowings.
5. Code-switching (CS): code-switching and grammatical calquing.

Specific annotation schemes are used for these tiers. The schemes for SyF, SeR, IST and also BOR-Phon tiers are versions of the corresponding schemes used in the Nganasan Spoken Language Corpus (NSLC) (Brykina et al., 2016) and documented in (Wagner-Nagy, Szeverényi, 2015), adapted to the INEL languages. The schemes for BOR, BOR-Morph and CS tiers have been developed specifically for the INEL corpora. The BOR tier combines information on the source language of borrowing and its lexical class (such as core vs. cultural vocabulary, grammatical markers, discourse markers and modal words). The CS tier distinguishes between several structural types of code-switching (sentence-external vs. sentence-internal, further subdivided). The same tier is used to mark grammatical calques (which may be only annotated on fragments in the main subject language and are thus mutually exclusive with code-switching).

The **metadata** for the analyzed materials include information about particular texts (communications) and participating speakers. Communication metadata comprise sections on content genre, involved speakers, data collection, processing and analysis ("who did what"), references to archive sources and publications (if any). Speaker metadata include sections on personal biography, ethnicity, family relations, education, known residence places, and language/dialect attribution.

This information, if available, is especially valuable in the context of intricate dialectal variation, language contacts and sporadic individual migrations, typical for much of the concerned geographical area.

The corpora will be gradually published starting from the end of 2018 through the technical infrastructure of the Hamburg Centre for Language Corpora (Hamburger Zentrum für Sprachkorpora – HZSK[3]). Thereby data is stored in XML-based the data format provided by the EXMARaLDA[4] software suite which guarantees a broad range of opportunities for the conversion, analysis and visualization of the data and thus sustainable long-term access. Registered users will be able to both download the annotated texts along with the basic metadata and examine the corpora with their local copy of the EXMARaLDA software and also perform web-based online search. Detailed user documentation is prepared and will be provided along with the corpora, cf. (Däbritz, 2017; Orlova et al., 2017).

Apart from corpora as the main outcome, **additional information resources** are being created to assist ongoing research. Since they may also be of interest to a wider academic community, their public versions are provided online. These include: a catalog of primary linguistic sources stored at various archives, such as written fieldwork materials and sound media (INEL Resource Catalog[5]); a bibliography of scholarly publications on the languages studied (INEL Bibliography[6]); geographic information on settlements where the studied languages are or were spoken (INEL Map[7]).

3. The Selkup corpus

The INEL Selkup corpus is composed of texts from the archive of Angelina Ivanovna Kuzmina (1924–2002), a student of Andrey P. Dulson and a renowned researcher of the Tomsk school. In 1962–1977, Kuzmina gathered a large amount of precious material on Selkup in almost all regions where the Selkup lived; this was a time when many Selkup varieties were still spoken. She collected both original and translated texts, grammatical and lexical data, as well as fragments of biographical and ethnographical nature (in Russian). In 2001, Kuzmina handed the archive over to Eugen Helimski (1950–2007), director of the Institute for Finno-Ugric/Uralic Studies at Hamburg University. The archive is since then preserved at IFUU; a detailed description was made by Eugen Helimski and Natalia Tuchkova (Tučkova, Helimski, 2010).

Kuzmina's archive includes both written and audio data. The written part comprises handwritten copybooks bound in 30 volumes (357 copybooks with 8554 pages). Some more Kuzmina's materials (mostly lexical and grammatical data) remain with her former student Lyudmila Ilyina in Novosibirsk, and some in the Dulson Archive in Tomsk (making up in sum another 5 volumes). The sound recordings, originally on over 30 magnetic tapes, were digitized in Novosibirsk by Galina Soldatova and are stored at IFUU in digital form. The written volumes contain a total of 295 texts: 79 folklore texts, 140 narratives, 25 translations from Russian; 51 translation from Northern Selkup to the Sondorovo dialect. Only a small fraction of them were published by Kuzmina herself (Kuzmina, 1967; 1974; 1977) or by Dulson (Dul'zon, 1966), some were published in recent years (Tučkova, Helimski, 2010; Tutschkova, Wagner-Nagy, 2015). These previous publications now make part of the Selkup Language Corpus (SLC, Budzisch et al., 2018). The INEL Selkup corpus, however, will be the first digital annotated publication of all these and other materials directly as represented in Kuzmina's archive.

[3] URL: https://corpora.uni-hamburg.de/hzsk/

[4] Extensible Markup Language for discourse Annotation: URL: http://exmaralda.org

[5] URL: http://corpora.uni-hamburg.de/inel/public-resource-catalogue/

[6] URL: https://inel.corpora.uni-hamburg.de/?page_id=1281

[7] URL: https://inel.corpora.uni-hamburg.de/?page_id=593

The digitized audio tapes contain ca. 85 texts, part of which have more or less exact correspondences in the manuscript archive; however, only in rare cases they can be taken as being proper transcriptions of the recording in question. It should also be noted that audio recording was used more extensively by Kuzmina in the Northern Selkup varieties, while the majority of the written notes belong to Central and Southern Selkup.

The work on Selkup is scheduled for two 3-year phases of the INEL project, with more focus on Northern varieties in the first phase (2016–2018) and on Central and Southern varieties in the second phase (2019–2021). The annotated corpus currently contains 43 of the 45 texts in Northern Selkup (1,904 sentences, 11,322 tokens); see below on trancriptions.

The core parts of the workflow are as described in the previous section. However, the preceding steps, i.e. transcription and translation, differ for three parts of the corpus. The majority of texts come from the manuscript part of the Kuzmina archive. These are typed in manually, respecting as close as possible Kuzmina's Cyrillic transcription. As noted in (Tučkova, Helimski, 2010: 14–15), the reliability of Kuzmina's transcription varies between and also within individual volumes of the archive, and transcripts of Northern Selkup are in general less precise. Quite often even the segmentation into words presents a significant problem, let alone the phonetic variation. Kuzmina used the transcription system developed by Andrey Dulson. Besides the 32 Cyrillic letters (except *ё*) it also recurs to some Latin, Greek and extended Cyrillic characters (*k, j, l, w, ɣ, ə, ɛ, ӊ*), additional diacritics (¨ ˚ ˜ ʿ ʾ ˛ ¯ ˆ ˳) and stress markers (´ `); one character may bear more than one diacritic. This original transcription is first automatically converted into Latin alphabet; then, in order to reduce variation and to make the glossing process easier, it is rendered into a near-phonemic transcription. Most of the changes are supposed either to merge allophones (e.g. in Northern Selkup voiced consonants are replaced with voiceless ones), or to correct Russian perception of the Selkup pronunciation (since the author of manuscripts was a native speaker of Russian), see in (1): *б → p, Г→ q, ä → ɛ.*

1. ′лыбыГың äса.
Lɨpɨ-qɨn ɛː-sa.
darkness-ADVLOC be-PST-[3SG.S]
'It was dark.'
[AR_1965_RestlessNight_transl.001]

Most of the archive texts are written down in Selkup with interlinear word-by-word translation into Russian. However, it is often fragmentary (only selected words are provided with a gloss) or altogether absent. Kuzmina's translation is therefore supplemented with a complete Russian free translation.

The audio recordings are being transcribed with the help of native speakers. In April 2017, a two-week working session was held in Moscow by Alexandre Arkhipov and Svetlana Orlova with two speakers of Northern varieties, Valentina Tamel'kina (Krasnosel'kup; Middle Taz dialect) and Svetlana Sankevich (Kunina) (Tarko-Sale; Upper Tol'ka dialect). The initial quality of many digitized tapes was rather poor, and some of them proved to be hardly intelligible even after sophisticated restoration procedures. However, 28 texts (with a total duration of 90 min, ca. 6,400 words) could have been transcribed during that session.

Some newly transcribed texts could be compared with matching parts in the manuscript archive (made by Kuzmina with the help of other speakers) and the latter proved to be quite different textually. Cf. the following versions of the same tale's beginning (Kuzmina's transcript in (2) and direct tape transcript in (3)):

2. Ili-mpɔː-tɨn kül'a-j n'oma-j əmä-si-n.
live-PST.NAR-3PL raven-ADJZ hare-ADJZ mother-CRC-PL.[NOM]
Ukkor č'onta-qɨn n'oma qən-pa šöt-tɨ.
one middle-LOC hare.[NOM] leave-PST.NAR.[3SG.S] forest-ILL
'There lived a raven hen and a doe hare with their children. Once the hare left for the forest.'
[KR_1969_RavensAndHares1_flk.002-003]

3. N'oma i kul'a ämɨ-sa-qaqı
hare.[NOM] and raven.[NOM] mother-CRC-DU.[NOM]
il'i-mpɔː-qı, il'i-mpɔː-qı.
live-PST.NAR-3DU.S live-PST.NAR-3DU.S
N'oma... n'oma šöt-tɨ qəla, šöt-tɨ
hare.[NOM] hare.[NOM] forest-ILL leave-CVB forest-ILL
qən-pa i š'ittalä…
leave-PST.NAR.[3SG.S] and then
'There lived a raven hen and a doe hare with their children. The hare left for the forest,
left for the forest and then…'
[KR_1969_RavensAndHares2_flk.001-002]

During the work with native speakers some questions on the previously analyzed written texts from Kuzmina's archive could be clarified, and some pilot phonetic recordings were made. The working sessions were also sound-recorded as a backup for informants' pronunciation and interpretation on later stages of analysis. Another similar fieldwork session is planned for the summer 2018.

Meanwhile, the work has already begun on transcribing Kuzmina's texts from Central and Southern Selkup dialects. Apart from digitizing texts from the manuscript collection, transcribing audio was also undertaken. In contrast to Northern Selkup, with two or three exceptions there are no speakers remaining for any of Central and Southern Selkup dialects who could work as consultants. Fortunately, the last semi-speaker of Narym dialect (Central), Irina Korobeynikova from Parabel was able to work remotely to transcribe sound recordings from Napas (Tym dialect, Central) and, assisted by Natalia Izhenbina, those from Ivankino on Ob (Middle Ob dialect, Southern), which is more distant from her own dialect. With their help, 18 texts of ca. 160 min total duration and 9,800 words are by now transcribed and translated into Russian. There still remains about 105 min of sound recorded in Ust'-Ozernoe (Ket dialect, Southern), for which it is not yet clear if transcribing by the same speakers could be comparably successful.

4. The Kamas corpus

The INEL Kamas corpus aims at bringing together the existing materials on the Kamas language (< Samoyedic < Uralic). Kamas is considered to be finally extinct since the death of its last speaker, Klavdiya Plotnikova, in 1989 (Klumpp, 2002: 27). The vitality of the language must have ceased much earlier, as already in 1925 the ethnographer Arkadiy Tugarinov could only find a handful of elders who still spoke the language (Tugarinov, 1926: 83). Kamas was documented by several researchers in XVIII–XX centuries. Major contributions were due to Matthias Alexander Castrén (in 1847) and Kai Donner (in 1912 and 1914) (Klumpp, 2002: 24f.). Castrén mostly collected lexical and grammatical material, parts of it published in 1854 and 1855. Parts of Kai Donner's collection – both lexical material and texts – were published posthumously by Aulis Joki (1944). During a fieldtrip on toponymics led by Aleksandr Matveyev in the village of Abalakovo in 1963, Klavdiya Plotnikova was identified as the last surviving Kamas speaker. Subsequently, Ago Künnap stepped in as the main researcher working with her, both in Abalakovo and in Estonia, producing a substantial amount of audio recordings and several publications of texts. At the same time, another Kamas rememberer, Aleksandra Semyonova, was found in Krasnoyarsk; unfortunately, she died soon afterwards.

The two main sources used in the INEL Kamas corpus are texts from Kai Donner's collection (a.k.a. "Pre-shift Kamas")[8] and audio recordings of Klavdiya Plotnikova made between 1964 and 1970 by Ago Künnap and other researchers (a.k.a. "Post-shift Kamas"). Texts collected by Kai Donner come from an unpublished edition of Kai Donner's manuscripts by Prof. Gerson Klumpp

[8] The terms "pre-shift (post-shift) Kamas" used by Gerson Klumpp refer to the state of Kamas as documented before (resp. after) the remnants of the Kamas community shifted entirely to Russian (see e.g. (Klumpp, 2013)).

(University of Tartu). The manuscripts were first philologically edited by Hartmut Katz and Gerson Klumpp. The resulting version was enhanced, rendered into a phonological transcription and morphologically glossed by Gerson Klumpp, and this edition was further adapted and edited within INEL. The German translation by A. Joki is taken from the published version (Joki, 1944), supplemented by a modern German version. The 16 texts include 12 folktales, 2 sets of riddles, a prayer and a song.

The audio material collected in 1964–1970 was made available by the Archive of Estonian Dialects and Kindred Languages (AEDKL) of the University of Tartu, Estonia, and by the KOTUS archive in Helsinki, Finland. The material stored in Tartu was collected by Ago Künnap and Tiit-Rein Viitso in Abalakovo and in Tartu, while Helsinki material was recorded by Jaakko Yli-Paavola in Tallinn (Klumpp, 2002: 27f.). These recordings represent a mixture of genres, alternating between folktales, ethnographic descriptions or personal stories, and translations of individual sentences from Russian stimuli. As for the folktales, many of them are evidently Russian, perhaps retold closely after a printed edition; some are retellings of Donner's tales from (Joki, 1944). Some texts are dialogical, other speakers addressing Plotnikova in Russian, with part of her replies also in Russian. Finally, there are two recordings of the other Kamas speaker, Aleksandra Semyonova, made in Krasnoyarsk around 1964.

The ultimate goal, within the current project, is to cover as much as possible of the audio from AEDKL and KOTUS. Another 20 tapes recorded by A. Matveyev are reported to exist in Ekaterinburg (Klumpp, 2013: 45 fn. 1), however these tapes are not yet digitized and cannot be studied in scope of the INEL project.

Transcribing and translating Kamas recordings is naturally problematic for the lack of living speakers to consult. However, the audio quality is overall quite good, and Plotnikova's low fluency in Kamas is also favorable in this respect, since her pronounciation is quite slow and careful. The phonological transcription used in the INEL Kamas corpus follows in most respects the transcription system developed by Gerson Klumpp, who has also provided extensive consultations during the project[9]. The tapes are highly fragmented, especially those containing elicited sentences (with up to 60–80 fragments per tape): the recording is stopped and restarted, omitting pauses but also most of researcher's prompts, which does not make understanding the contents any easier. As for the retellings of Russian tales, they are also hard to interpret unless the original story can be identified; the same anaphoric element, *dĭ*, is used for all participants, and many logical connections are missing. Fortunately, for many of the tales close Russian equivalents could be found.

The language of the two parts of the corpus differs to a great extent. Plotnikova's Kamas is heavily eroded and shows Russian influence on all levels. Morphology and syntax are very limited, with incomplete paradigms and hardly any constructions more complex than direct speech, infinitive or converb clauses.

Texts from Donner's collection contain fewer traces of direct Russian influence, while Plotnikova's tapes exhibit a lot of loanwords, code-switching and grammatical calques (cf. the use of dative in (5)):

4. ***Na što*** *(tagaj-ziʔ)* *tagaj* *i-bie-lʔ?*
for.what[RUS] knife-INS knife take-PST-2SG?
'What for did you take the knife?'

5. *Tănan* ***iššo*** ***nada****...*
you.DAT also[RUS] should[RUS]
'You also have to…'
[PKZ_1964_SU0205.174–175]

[9] Parts of transcriptions were done by Tiina Klooster in cooperation with Gerson Klumpp, and partial transcriptions from their earlier project served as basis for transcriptions of corresponding fragments in INEL.

At present, the annotated Kamas corpus contains 16 texts from Donner's collection (ca. 2,500 tokens) and 12 completely glossed Plotnikova's tapes (ca. 3,000 sentences and 10,500 tokens), another 13 tapes are transcribed and translated entirely and 6 partially.

5. The Dolgan corpus

The INEL corpus of Dolgan (< North-Siberian Turkic < Siberian/North-Eastern Turkic < Turkic) aims to collect, digitize and annotate Dolgan data from various sources. Despite of the fact that Dolgan is still spoken by approx. 1,000 people (VPN 2010) on the Taimyr Peninsula and in adjacent areas, it has to be regarded as a highly endangered language (cf. Siegl, 2013). The existing descriptions of Dolgan (Ubryatova, 1985; Li, 2011; Artem´ev, 2013) vary significantly in scope and quality, thus the language calls for further corpus-based research.

The Dolgan corpus at hand consists roughly of three parts: 1) previously published folklore texts (Efremov et al., 2000), 2) audio materials made available by the House of the Cultures of Taimyr Peninsula in Dudinka (henceforth: TDNT) and 3) fieldwork material from Eugénie Stapert. All material is either transliterated or transcribed using a latin-based transcription system. The core steps in processing the data are the same as described above. Besides that, part of the Dolgan data is also annotated for information structure relations, i.e. for topic-comment and focus-background structure. The scheme and principles for the annotation of information structure were especially developed for the project, based on the Leipzig Model of information structure (cf. e.g. Junghanns, Zybatow, 2009). In what follows, the material of the corpus will be described in more detail.

The texts published in (Efremov et al., 2000) were collected by various researchers (among them A. A. Popov and E. I. Ubryatova) in the 1930s and 1960s/1970s on the Taimyr Peninsula. Being published in a folklore volume, all texts can be subsumed under the broad "folklore" genre, including tales, legends and myths. All in all, 35 texts from this volume are used in the corpus, containing 3,206 sentences with 19,494 tokens. The language of these folklore texts has to be commented on shortly. During fieldwork in summer 2017 it became clear that native speakers of Dolgan tend to reject these texts as unnatural and "Yakut-like" while working on them. Hence, care should be taken concerning those texts; unfortunately, one cannot exclude the possibility that they underwent more or less substantial post-editing, which probably made them closer to literary Yakut. Nevertheless, the texts compile valuable material on Dolgan from both a linguistic and an ethnographic/folkloristic position.

The second and presumably the biggest part of the corpus is made up of audio materials which were made available by the TDNT. Their overall duration approaches 20 hours, of which ca. 7,5 hours have been transcribed to date. The recordings contain monologues (both folklore and narrative texts) and conversations (interviews), as well as a few songs; some texts are translations from Russian, while some others represent recited literary work by Dolgan writers. By now, the glossed corpus contains 24 texts with 2,183 sentences and 16,263 tokens from these recordings (corresponding to ca. 3 hours duration).

The transcription and translation into Russian is done by native speakers of Dolgan in Dudinka under the direction of Nina Kudryakova, Head of the department of folklore and ethnography of the TDNT. Having worked for a long time at the local radio and also as editor of folklore publications, she is doing an excellent job in transcribing, translating and editing the transcripts. It should be emphasized that without the effort and commitment of Nina Kudryakova and other Dolgan speakers the corpus would be much smaller than it is now.

The third part of the corpus stems from fieldwork material collected by Eugénie Stapert (University of Leiden) on fieldwork trips in 2008, 2009 and 2010. All in all, this material contains 7,5 hours of audio recordings, which were transcribed by Eugénie Stapert and local consultants partly in 2008–2010 and partly on a fieldwork trip to Dudinka in 2017 (see below). Approximately half of the material is by now transcribed and is being glossed. The texts are mostly monologic and are mostly narratives, however, there are some songs, too.

In summer 2017 Eugénie Stapert as visiting Fellow at the University of Hamburg and Chris Lasse Däbritz had the opportunity to conduct intensive fieldwork with Dolgan consultants in Dudinka. The goal of the fieldwork was mostly to answer open questions and to work with the material collected previously. Therefore, no new recordings were made specifically for the INEL corpus. The four weeks of fieldwork pushed the project significantly forward and gave good insights into both language and culture of the Dolgan community on the Taymyr peninsula.

As for the language of the the corpus, it has to be mentioned that the language of the published folklore tales is much more straightforward and clear which is no surprise as it was presumably edited. The audio recordings (both from TDNT and from fieldwork sessions), on the other hand, show much more natural discourse phenomena such as pauses, false starts, self-repairs, code-switching etc. Compare (6) and (7) to get an impression of the difference between the kinds of the texts:

6. *Bu* *emeːksin-iŋ* *bu* *hahɨl-ɨ* *ild'-en*
this old.woman-2SG.POSS this fox-ACC carry-CVB.SEQ
hül-ün-n-e, *bu* *hül-en* *baran* *kuːr-t-a.*
skin-MED-PST1-3SG this skin-CVB.SEQ after dry-PST1-3SG
'This old woman carried the fox and skinned it, after the skinning it dried.'
[FeA_1931_OldWomanFoxFur_flk.022; published folklore text]

7. *Kalxoːz-ka* *üleliː-r* *er-dek-pine,* *oččogo*
kolkhoz-DAT/LOC work-PTCP.PRS be-TEMP-1SG then
Bol'šakov G'eorg'ij… *Gavr'il,* *otčestva-tɨ-n*
Bolshakov.Georgiy Gavril patronymic-PX3SG-ACC
umnu-bup-pun, *pr'eds'edat'el'em kalxoza*
forget-PST2-1SG [as.director.of.the.kolkhoz.RUS]
ülel-eːčči *e-t-e.*
work-PTCP.HAB be-PST1-3SG
'As I was working at the kolkhoz, then Bolshakov Georgiy… Gavril, I forgot his patronymic,
he was working as director of the kolkhoz.'
[AkNN_KuNS_200212_LifeHandicraft_conv.015-016; radio interview]

Such natural language as shown in (7) and dialogues in general are obviously a gain for any corpus. Hence, the Dolgan material in the corpus is diverse and allows for research on a broad range of topics, including issues of pragmatics and contact phenomena. By now, the corpus contains 59 fully glossed and partly annotated texts with 5,309 sentences and 35,757 tokens.

6. Conclusion

The three corpora outlined above, Kamas, Selkup and Dolgan, will be the first bricks of the research infrastructure developed within the INEL project. Each of these presents unique material which can be fruitfully researched on its own. However, their combined value can and will be made even greater through development of common interfaces, integration with similar projects hosted at HZSK (such as the Nganasan Spoken Language Corpus and the Selkup Language Corpus), as well as additional information resources such as bibliographies, catalogs and maps.

References:

Artem'ev N. M. Dolganskij yazy`k. 10–11 klassy`. Uchebnoe posobie dlya obshcheobrazovatel`ny`x uchrezhdenij [Dolgan language. Textbook for secondary school]. – Sankt-Peterburg: Almaz-Graf, 2013. (in Russian)

Brykina Maria, Gusev Valentin, Szeverényi Sandor, Wagner-Nagy Beáta. Nganasan Spoken Language Corpus (NSLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 12.06.2018. URL: http://hdl.handle.net/11022/0000-0007-C6F2-8
Budzisch Josefina, Harder Anja, Wagner-Nagy Beáta. Selkup Language Corpus (SLC). – 2018. – Unpublished.
Comrie Bernard, Haspelmath Martin, Bickel Balthasar. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. – 2008. – Online at: URL: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf [last access: 17.02.2018].
Däbritz Chris Lasse. INEL Dolgan corpus. User documentation. – 2017. – Unpublished manuscript.
Dul'zon A. P. Ketskie skazki [Ket fairytales]. – Tomsk: Izdatel`stvo Tomskogo universiteta, 1966. – P. 128–150 (even), P. 137–155 (odd). (in Russian)
Efremov P. E. et al. (eds.). Fol`klor Dolgan [Folklore of the Dolgans]. Pamyatniki fol`klora narodov Sibiri i Dal`nego Vostoka 19. – Novosibirsk: Izdatel`stvo Instituta Arxeologii i E`tnografii SO RAN, 2000. (in Russian)
Joki Aulis J. Kai Donners Kamassisches Wörterbuch nebst Sprachproben und Hauptzügen der Grammatik. Lexica Societatis Fenno-Ugricae VIII. – Helsinki: Suomalais-Ugrilainen Seura, 1944.
Junghanns Uwe, Zybatow Gerhild. Grammatik und Informationsstruktur // Gutschmidt, Karl et al. (eds.). Die slavischen Sprachen. Handbücher zur Sprach- und Kommunikationswissenschaft. Bd. 32, 2. – Berlin: De Gruyter, 2009. – S. 684–707.
Klumpp Gerson. Konverbkonstruktionen im Kamassischen. Veröffentlichungen der Societas Uralo-Altaica 58. – Wiesbaden: Harrassowitz, 2002.
Klumpp Gerson. On Kai Donner's phonograph records of Kamas. Finnisch-Ugrische Mitteilungen 37. – 2013. – S. 45–59.
Kuz'mina A. I. Dialektologicheskie materialy` po sel`kupskomu yazy`ku. [Dialectological materials on Selkup] // Issledovaniya po yazy`ku i fol`kloru 2. – Novosibirsk: Nauka, 1967. – P. 267–329. (in Russian)
Kuz'mina A. I. Grammatika sel`kupskogo yazy`ka [A grammar of Selkup]. – Novosibirsk, 1974. – P. 149–150. (in Russian)
Kuz'mina A. I. K e`timologii nazvanij mesyacev, storon sveta, zvyozd i sozvezdij v sel`kupskom yazy`ke [Towards the etymology of the names of months, cardinal directions, stars and constellations in Selkup]. In: Yazy`ki i toponimiya 4. – Tomsk, 1977. – P. 71–85. (in Russian)
Li Yong-Song. A study of Dolgan. Altaic Languages Series 5. – Seoul: Seoul National University Press, 2011.
Siegl Florian. The Sociolinguistic status quo on the Taimyr Peninsula. Études finno-ougriennes [En ligne] 45. – 2013. Online at: URL: http://journals.openedition.org/efo/2472, [last access: 07.02.2018].
Tugarinov A. Ya. Poslednie kalmazhi [The last Kalmazhi] // Severnaya Aziya. – 1926. – Vol. 1. – P. 73–88.
Tučkova Natalia, Helimski Eugen. Über die selkupischen Sprachmaterialien von Angelina I. Kuzmina. Hamburger Sibirische und Finno-Ugrische Materialien, 5. – Hamburg, 2010.
Tutschkova Natalia, Wagner-Nagy Beáta. Texte über Itte, der über die sieben Götter der Weisheit verfügt. – Tomsk, 2015. (in Russian)
Orlova Svetlana, Brykina Maria, Arkhipov Alexandre. INEL Selkup corpus. User documentation. – 2017. – Unpublished.
Ubryatova Elizaveta I. Yazyk noril`skix dolgan. – Novosibirsk: Nauka, 1985. (in Russian)
VPN 2010 = Vserossijskaja perepis' naselenija. Tom 4. Nacional'nyj sostav i vladenie jazykami, graždanstvo. Online at: URL: http://www.gks.ru/free_doc/ new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf, [last access: 07.02.2018].
Wagner-Nagy Beáta, Szeverényi Sándor, Gusev Valentin. User's Guide to Nganasan Spoken Language Corpus // Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology. – 2018. – Vol. 1, no. 1. – P. 1–45. Online at: URL: http://www.iskolakultura.hu/index.php/wpcl/article/view/10611/10503, [last access: 17.06.2018].

Arkhipov A.
Universität Hamburg, Institut für Finnougristik/Uralistik.
Research fellow.
Max-Brauer-Allee, 60, Raum 019, D-22765 Hamburg.
Lomonosov Moscow State University, Institute of the World Culture.
Head of department of linguistic and cultural ecology, candidate of philological sciences.
Leninskie gory, 1, Moscow, Russia, 119991.
E-mail: alexandre.arkhipov@uni-hamburg.de, sarkipo@mail.ru

Däbritz Chris Lasse, research fellow. M.A.
Universität Hamburg, Institut für Finnougristik/Uralistik.
Überseering 35, D-22297 Hamburg.
E-mail: chris.lasse.daebritz@uni-hamburg.de

Материал поступил в редакцию 20 апреля 2018 г.

А. В. Архипов, К. Л. Дэбритц

ГАМБУРГСКИЕ КОРПУСА ЯЗЫКОВ НАРОДОВ СЕВЕРНОЙ ЕВРАЗИИ

В рамках долгосрочного исследовательского проекта INEL (2016–2033) в Гамбургском университете разрабатываются электронные корпуса текстов и сопутствующая инфраструктура для ряда языков Северной Евразии. В настоящий момент идёт работа по созданию корпусов селькупского, камасинского и долганского языков. В проекте используются архивные материалы из разных источников, в т. ч. хранящийся в Гамбурге селькупский архив А. И. Кузьминой, камасинские аудиозаписи из архивов в Тарту и Хельсинки, и долганские записи, предоставленные Таймырским домом народного творчества. Все тексты в корпусах снабжаются фонологической транскрипцией, поморфемным глоссированием, переводами; в части текстов также размечаются отдельные семантические и синтаксические признаки, информационный статус именных групп, заимствования и переключения кодов. Корпуса предназначены в первую очередь для проведения типологически ориентированных исследований в области грамматики, но могут представлять и более широкий интерес. Кроме того, для повышения эффективности лингвистических изысканий разрабатывается ряд вспомогательных информационных ресурсов.

Ключевые слова: *проект INEL, корпуса, селькупский язык, камасинский язык, долганский язык, архивные материалы, грамматика, языковые контакты, поморфемное глоссирование, разметка.*

Архипов А. В.
Гамбургский университет, Институт финно-угроведения/уралистики.
Научный сотрудник.
Макс-Брауэр-Аллее 60, R. 019, D-22765 Hamburg.
МГУ им. М. В. Ломоносова, Институт мировой культуры.
Ленинские горы, 1, Москва, Россия, 119991.
Зав. отделом лингвокультурной экологии, кандидат филологических наук.
E-mail: alexandre.arkhipov@uni-hamburg.de, sarkipo@mail.ru

Дэбритц Крис Лассе.
Гамбургский университет, Институт финно-угроведения/уралистики.
Юберзееринг 35, D-22297, Hamburg.
Научный сотрудник. M.A.
E-mail: chris.lasse.daebritz@uni-hamburg.de