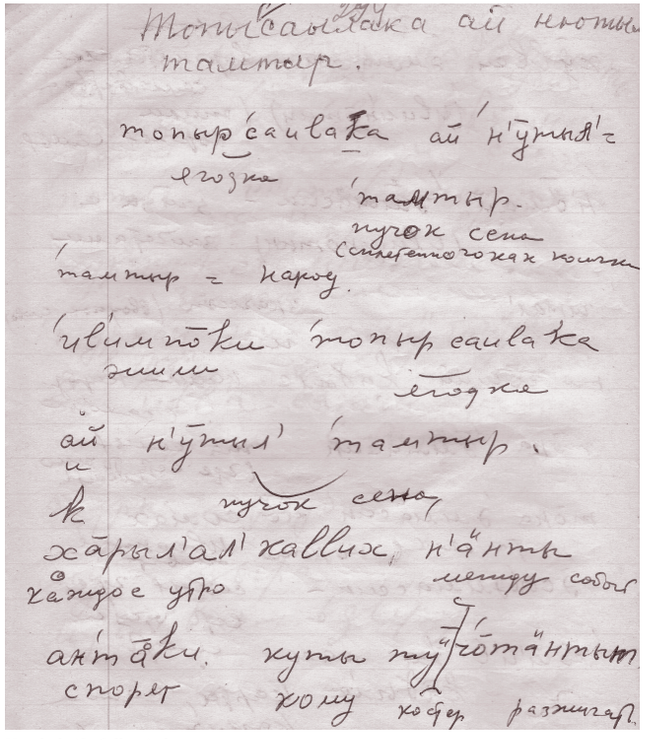
Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages

Duration
01/02/2016–31/12/2033

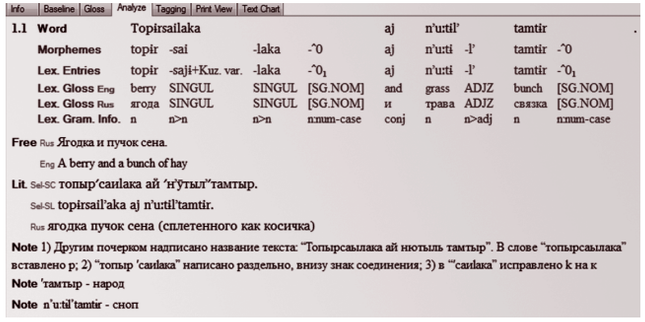
Goal
Providing language resources for indigenous languages and creating a digital research infrastructure for the use of these resources.

Project description
The territory of the Russian Federation is home to a wide variety of indigenous languages, of which some are spoken actively by only a few inhabitants of the region and others are acutely endangered, if not extinct.
The existence of these languages has been documented in numerous materials, including handwritten texts, analog sound recordings, word lists, etc., which are currently located in archives and collections (e.g. in Hamburg, Helsinki, Moscow, St. Petersburg, Tartu and Tomsk).



Handschrift, Selkupisch
Manuscript, Selkup

The discovery and development of these resources – in some cases even the documentation of their existence – is one of the initial activities that INEL has been pursuing since the beginning of 2016. Using the latest methods of digital data processing, the unique data contained in the collections are being collected, partially scaled, digitized and enriched with a wealth of other linguistic information.



Info	Baseline	Gloss	Analyze	Tagging	Print View	Text Chart

1.1	Word	Topirsailaka				aj	n'u:til'		tamtir	
	Morphemes	topir	-sai	-laka	-^0	aj	n'u:ti	-l'	tamtir	-^0
	Lex. Entries	topir	-saji+Kuz. var.	-laka	-^0$_1$	aj	n'u:ti	-l'	tamtir	-^0$_1$
	Lex. Gloss Eng	berry	SINGUL	SINGUL	[SG.NOM]	and	grass	ADJZ	bunch	[SG.NOM]
	Lex. Gloss Rus	ягода	SINGUL	SINGUL	[SG.NOM]	и	трава	ADJZ	связка	[SG.NOM]
	Lex. Gram. Info.	n	n>n	n>n	n:num-case	conj	n	n>adj	n	n:num-case

Free Rus Ягодка и пучок сена.
Eng A berry and a bunch of hay
Lit. Sel-SC топыр'саилака ай 'н'ӱтыл"тамтыр.
Sel-SL topirsail'aka aj n'u:til'tamtir.
Rus ягодка пучок сена (сплетенного как косичка)
Note 1) Другим почерком надписано название текста: "Топырсаылака ай нютыль тамтыр". В слове "топырсаылака" вставлено р; 2) "топыр 'саилака" написано раздельно, внизу знак соединения; 3) в "'саилака" исправлено k на к
Note 'гамтыр - народ
Note n'u:til'tamtir - сноп

Digitalisierte und mit linguistischen Informationen angereicherte Version der Handschrift
Digitized and linguistically annotated version of the manuscript

The resulting digital empirical data collections (corpora) will be made permanently available online in an interdisciplinary network. INEL acts as a bridge between the exploration of indigenous languages and the international partner institutions.

Cooperation partner
University of Hamburg

Principal Investigator
Prof. Dr. Beáta Wagner-Nagy

Financing
The long-term project is funded within the framework of the Academies' Programme, which is coordinated by the Union of the German Academies of Sciences and Humanities.

UNION
DER DEUTSCHEN AKADEMIEN
DER WISSENSCHAFTEN

AKADEMIE DER
WISSENSCHAFTEN
IN HAMBURG

Forschungsprojekt im Akademienprogramm

INEL

Grammatiken, Korpora und Sprachtechnologie für indigene nordeurasische Sprachen

Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages

www.awhamburg.de

Nenzisches Rentierfest (1)
Nenets reindeer festival (1)

Grammatiken, Korpora und Sprachtechnologie für indigene nordeurasische Sprachen

Laufzeit
01.02.2016 – 31.12.2033

Zielsetzung
Erschließung sprachlicher Ressourcen indigener Sprachen sowie ihre Bereitstellung über eine digitale Forschungsinfrastruktur.

Projektbeschreibung
Auf dem Gebiet der Russischen Föderation existiert eine Vielfalt indigener Sprachen, die lediglich von wenigen Bewohnern der Region noch aktiv gesprochen werden und zum Teil sogar akut bedroht oder ausgestorben sind.
Die zahlreichen Zeugnisse hierüber liegen beispielsweise in Form von handgeschriebenen Textsammlungen, analogen Tonaufnahmen, Wortlisten etc. in Archiven und Sammlungen (z. B. in Hamburg, Helsinki, Moskau, St. Petersburg, Tartu, Tomsk, etc.) vor.

Das Auffinden und die Erschließung dieser Ressourcen – in einigen Fällen überhaupt die Dokumentation ihrer Existenz – ist eine der initialen Tätigkeiten, die INEL seit Beginn des Jahres 2016 verfolgt. Mithilfe aktueller Methoden der digitalen Datenaufbereitung werden die in den Sammlungen enthaltenen einzigartigen Daten zusammengetragen, teilweise nacherhoben, digitalisiert und mit einer Fülle weiterer linguistischer Informationen angereichert.
Die daraus entstehenden digitalen empirischen Datensammlungen (Korpora) werden in einem interdisziplinären Netzwerk der Öffentlichkeit dauerhaft online bereitgestellt.
Dabei fungiert INEL als Brücke zwischen der Erforschung indigener Sprachen und den internationalen Partnerinstitutionen.

Die im INEL-Projekt behandelten Sprachen stammen aus den Sprachfamilien Uralisch (Selkupisch, Kamassisch, Nenzisch und Komi) und Altaisch (Dolganisch, Ewenkisch und Sibirisch-Tatarisch). Darüber hinaus wird die isolierte Sprache Ketisch behandelt. (2)
The languages dealt with in the INEL project are from the Uralic (Selkup, Kamas, Nenets and Komi) and Altaic (Dolgan, Evenki and Siberian Tatar) families. Also included is Ket, an isolate language. (2)

Kooperationspartner
Universität Hamburg

Leitung
Prof. Dr. Beáta Wagner-Nagy

Finanzierung
Das Langzeitvorhaben wird im Rahmen des Akademienprogramms gefördert, das von der Union der deutschen Akademien der Wissenschaften koordiniert wird.

UNION
DER DEUTSCHEN AKADEMIEN
DER WISSENSCHAFTEN

AKADEMIE DER
WISSENSCHAFTEN
IN HAMBURG

Kontakt

INEL Projekt
Institut für Finnougristik/Uralistik
Universität Hamburg
Max-Brauer-Allee 60
22765 Hamburg

Ansprechpartner:
Dr. Alexandre Arkhipov

inel@uni-hamburg.de
Tel.: +49 40 42838 6890
http://inel.corpora.uni-hamburg.de

INEL

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Stand: November 2016
Herausgeber: Präsident der Akademie der Wissenschaften in Hamburg
Grafik: Hubert Eckl, KommunikationsDesign
Abbildungen: Aleksej Soin (1), Karte von Openstreetmap unter der Lizenz CC BY-NC-SA 2.0 (2), Kuzmina Archiv, Hamburg
Druck: Bartels Druck GmbH