The INEL Dolgan corpus: Insights into an endangered language of Northern Eurasia[1]

Chris Lasse Däbritz

The paper at hand presents a description of the INEL Dolgan Corpus that has been created from 2016 to 2019 within the INEL project, located at the Institute for Finno-Ugric/Uralic Studies of the University of Hamburg. The corpus aims to provide a digital research infrastructure for Dolgan, an indigenous language of Northern Siberia. Though Dolgan is a Turkic language, the corpus is relevant for researchers of Uralic languages both due to the close areal connections of Uralic with Dolgan on the Taymyr peninsula and on account of the fact that it is an example of electronic research infrastructure developed for an endangered language. After introducing Dolgan and the INEL project, the paper describes the INEL Dolgan Corpus in detail, focusing on its linguistic content, annotation layers and search possibilities. Finally, the paper provides an outlook on how the corpus contributes to furthering research on this endangered language.

Keywords: *corpus, INEL project, Dolgan, languages of Northern Siberia, endangered languages*

1 Introduction

Dolgan is a Turkic language that is spoken by 1,054 people (VPN 2010) on the Taymyr peninsula and in adjacent areas in the extreme north of the Russian Federation. Several features call for the documentation and investigation of this indigenous language of Northern Siberia. First, Dolgan has been regarded a dialect of Sakha (Yakut) for a long time. As recently as in the 1980s, Ubrjatova (1985) pointed out that Dolgan is a language on its own that arose from Sakha (Yakut) under heavy Evenki (< Tungusic) substrate. Until today this has led to many accounts to Dolgan that are biased by Sakha (Yakut). Second, Dolgan was and is in contact with many surrounding languages (Sakha (Yakut), Evenki, to a lesser extent Nganasan and Enets, as well as Standard Russian, local Russian varieties and Taymyr Pidgin Russian). Especially the contact scenario, out of which Dolgan arose, is not fully understood yet, neither is the intensity of possible Samoyedic–Dolgan contacts. Therefore, the investigation of Dolgan has a particular relevance for Samoyedic studies, too. Finally – like many other indigenous languages of Siberia – Dolgan faces extinction, which is a sufficient reason on its own for conducting documentation work, collecting language material and compiling a linguistic corpus.

The INEL Dolgan Corpus[2] aims at founding the empirical base for the investigation of the language, which is the main goal of all INEL corpora (see section 2). In order to reach this goal, material from as many sources as possible is collected, digitized and linguistically annotated; moreover, some linguistic research already has been carried out on the basis of the INEL Dolgan Corpus (see section 3). Finally, the INEL Dolgan Corpus may, thus, contribute to an up-to-date documentation of Siberian languages, being useful for a wide range of both linguistic and even non-linguistic research (see section 4).

[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] PID: http://hdl.handle.net/11022/0000-0007-CAE7-1

ISSN: 2063-8825

2 INEL and the INEL corpora

The acronym "INEL" stands for *Grammatical Descriptions, Corpora and Language Technology for **I**ndigenous **N**orthern **E**urasian **L**anguages*, and refers to a long-term research project, being carried out at the Institute for Finno-Ugric/Uralic Studies of the University of Hamburg.[3] Its major aim is to create digital linguistic corpora as well as research infrastructure for several lesser-described Northern Eurasian languages and varieties. It is scheduled for 18 years (2016–2033), allowing three years for each language/variety dealt with. Table 1 shows the finalized and ongoing subprojects. In the future, further languages such as Ket and Nenets (Taymyr and Kanin variety) are planned to be included.

Language	Period
Selkup (all varieties)	01/2016 – 12/2021
Kamas	01/2016 – 12/2018
Dolgan	09/2016 – 08/2019
Evenki (Northern and Southern varieties)	01/2019 – 12/2021

Table 1: *Languages dealt with in the INEL project*

As can be seen from the table above, the languages dealt with in the INEL project come mostly from Western Siberia, being under-resourced and exhibiting clear areal connections. Although the INEL project contributes to the documentation of these languages, it differs from many language documentation projects in an important way: The material that is processed often comes from existing archives and collections, rather than being collected within the project itself. This leads to a broad variety of material included, which will be described in detail for Dolgan in section 3. This language material is digitized and, thus, made accessible for linguistic annotation and the compilation of linguistic corpora. Up to now, the INEL project published three open-access corpora, namely the *INEL Selkup Corpus* (Brykina et al. 2020), the *INEL Kamas Corpus* (Gusev et al. 2019), and the *INEL Dolgan Corpus* (Däbritz et al. 2019).[4] The following Table 2 sums up basic statistical information on those corpora.

[3] The principal investigator is Prof. Beáta Wagner-Nagy, and the funding was applied for by Prof. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland and Timm Lehmberg. The current project members are Prof. Dr. Beáta Wagner-Nagy, Dr. Alexandre Arkhipov (research coordinator), Timm Lehmberg (technical coordinator), Dr. Maria Brykina, Chris Lasse Däbritz (linguistic team), Anne Ferger, Daniel Jettka (technical team). The project website is available at https://www.slm.uni-hamburg.de/inel/.

[4] The corpora are available under the terms and conditions of Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License (CC BY-NC-SA 4.0), cf. https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode, last access: 22/04/2020.

Corpus	Transcripts[5]	Tokens	Speakers	Genres
INEL Selkup	264	42,466	74	folklore, narrative, translations, songs, conversations
INEL Kamas	158	63,824	4 (+ 2 unknown)	folklore, narrative, songs, miscellaneous (e.g. riddles)
INEL Dolgan	116	77,636	61	folklore, narrative, translations, songs, conversations

Table 2: *INEL corpora – statistics*

All INEL corpora are compiled following similar principles and guidelines. However, each corpus certainly has its peculiarities and special characteristics. The INEL Selkup Corpus is composed of the personal archive of Angelina Ivanova Kuzmina (1924–2002). It explicitly aims at covering all dialects of Selkup, which makes possible comparative studies of Northern, Central and Southern dialects. The INEL Kamas Corpus – as can be seen from the table – has a much smaller amount of speakers included, which is of course to be explained by the fact that Kamas is extinct, and there is simply no more material available. Nevertheless, the corpus contains transcripts from a relatively wide range of time, including both old texts from the 1910s and newer texts of Klavdiya Plotnikova from the 1960s and 1970s. The INEL Dolgan Corpus, finally, is the first corpus that covers a language, which is to some extent spoken in everyday life. Therefore, it was possible to include a higher amount of free conversations (radio interviews) into the corpus than in the cases of Selkup (especially Central and Southern dialects) and Kamas.

Thus, the INEL project provides an infrastructure for the compilation of structurally similar corpora of diverse languages, including diverse language material. For a concise description of the INEL project in general as well as those corpora, see also Arkhipov & Däbritz (2018).

3 The INEL Dolgan Corpus

As was mentioned already in the introduction, the INEL Dolgan Corpus aims at enabling the investigation of this rarely studied indigenous language of Northern Siberia on an empirically solid base. Given this, the content of the INEL Dolgan Corpus has to fulfil several criteria: as balanced a provenance as possible, as transparent a linguistic representation as possible and as accessible a technical representation as possible. The following paragraphs describe how the INEL Dolgan Corpus seeks to fulfil these criteria. The material included into the INEL Dolgan corpus comes from four very different sources:

1) texts from the published volume *Fol'klor Dolgan* [FD 2000] (Efremov et al. 2000),
2) audio material obtained from the *Taymyr House of National Arts* (TDNT),
3) audio material obtained from the collection of Eugénie Stapert, and
4) audio material collected during fieldwork in Dudinka in 2017.

[5] The term “transcript” is used here as a cover term for all items (texts, conversations or the like) included into the corpus.

Overall, the INEL Dolgan Corpus contains 116 transcripts (16 conversations, 50 folklore texts, 44 narratives, 2 songs, 4 translations from Russian) of 61 speakers (33 female, 28 male) with 11,329 utterances and 77,636 tokens. 81 communications can be linked to a corresponding audio file, making up a total of 10:42:14 hours of audio material. The following Figure 1 shows the number of tokens (green bars) and communications (blue bars) of each genre.

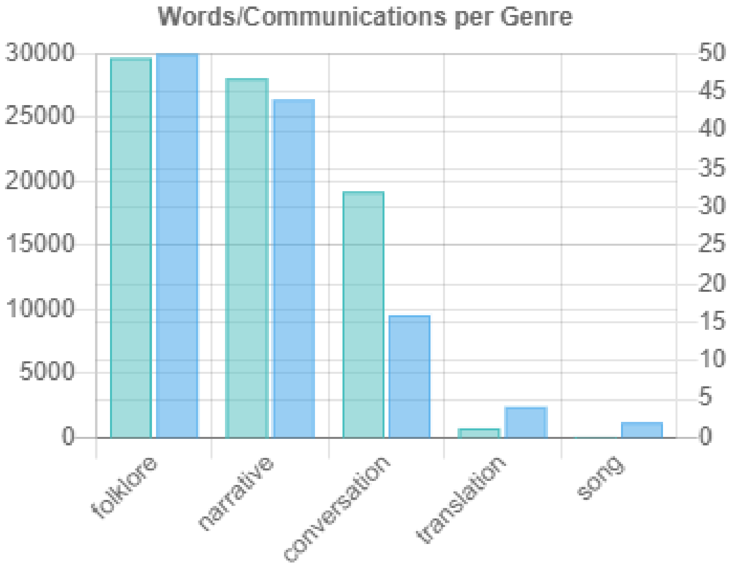


Figure 1: *Words/communications per genre in the INEL Dolgan Corpus*

The INEL Dolgan Corpus is published through the INEL infrastructure, the latter being partly based on existing infrastructure of the Hamburg Center for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK).[6] The data is stored in XML-based format provided by the EXMARaLDA program package.[7] To be able to browse the corpus and use the data locally, the relevant software tools (Partitur Editor[8], Corpus Manager[9], EXAKT[10]) have to be installed. In addition, the corpus – like the other INEL corpora, too – can be searched online using the Tsakonian Corpus Platform[11] (see Arkhangelskiy, Ferger & Hedeland 2019 for technical details).

As for the content of the communications, there is always a phonological transcription of the Dolgan speech, morphological glossing as well as further annotations and translations into various languages. The principles of transcribing, glossing, annotating and translating are summarized in a user documentation file that is provided with the corpus data[12], and is additionally published (Däbritz 2020).

The phonological transcription is based on principles used in all INEL corpora, which include elements from both IPA and FUT, the morphological glossing follows the

[6] http://hdl.handle.net/11022/0000-0007-CAE7-1, last access: 27/04/2020

[7] https://exmaralda.org/en/, last access: 27/04/2020

[8] https://exmaralda.org/en/partitur-editor-en/, last access: 27/04/2020

[9] https://exmaralda.org/en/corpus-manager-en/, last access: 27/04/2020

[10] https://exmaralda.org/en/exakt-en/, last access: 27/04/2020

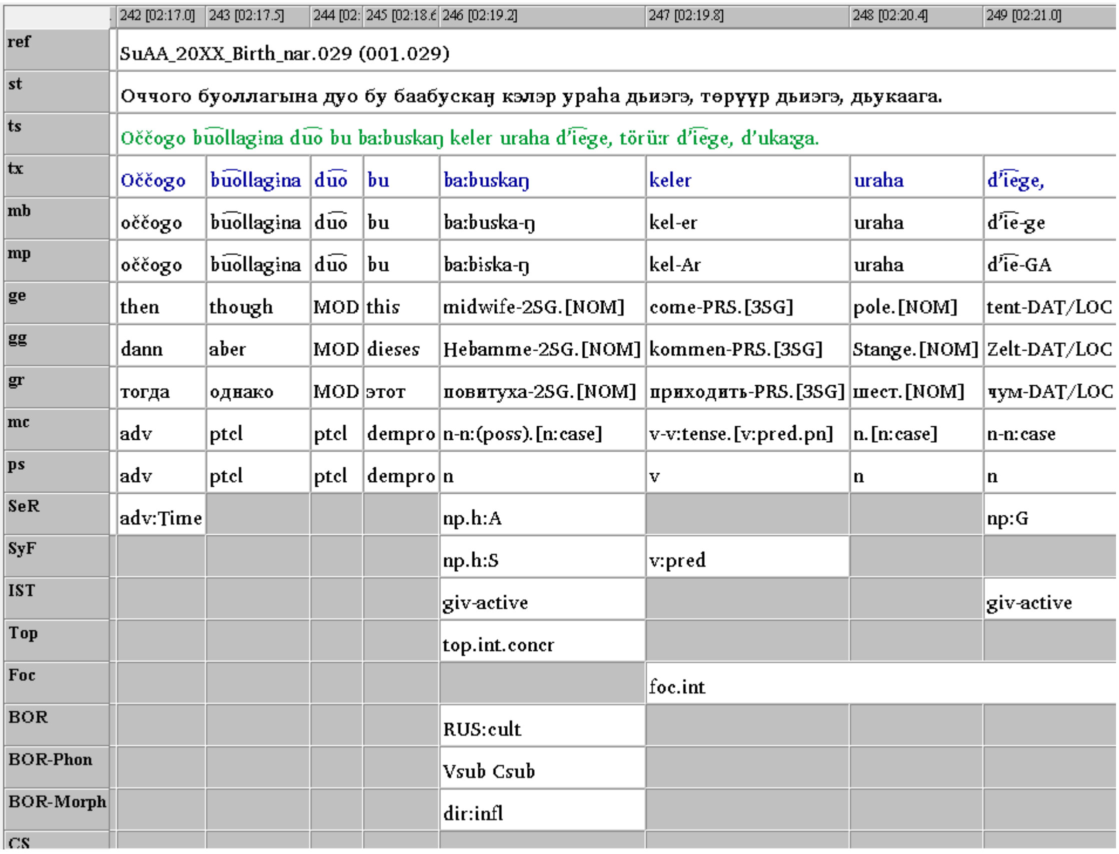
[11] https://bitbucket.org/tsakorpus/, last access: 27/04/2020. Search can be performed through the following link: https://inel.corpora.uni-hamburg.de/DolganCorpus/search

[12] http://hdl.handle.net/11022/0000-0007-CAE7-1, last access: 28/04/2020.

principles of the Leipzig Glossing Rules (2015).[13] Lexical glosses are provided in English, German and Russian; grammatical glosses do not differ between the languages of analysis. Further annotation tiers contain the annotation of Semantic Roles (SeR), Syntactic Functions (SyF), Information Status (IST), Information Structure (Top and Foc), Borrowing (BOR) and Code-switching (CS). The annotations of SeR, SyF and IST are based on the principles developed for the *Nganasan Spoken Language Corpus* (NSLC; Brykina et al. 2018), described by Wagner-Nagy et al. (2018). The annotations of Top, Foc, BOR and CS were developed within the INEL project in close cooperation with the compilers of NSLC, see also Arkhipov (2020) for details of the latter two. Finally, free translations into English, German and Russian are provided. If the transcript was already published (transcripts from FD 2000) or had been translated by our native speaker assistants (transcripts from TDNT), this literal translation is given, too.

The deep annotation of the corpus data enables the user to conduct varied and complex searches. The grammatical glossing is form-oriented, i.e. grammatical forms are analyzed with respect to their components. As an example, Figure 2 contains the item *babuska-ŋ* 'midwife-2SG', which would be found via a search of *midwife* or the possessive suffix of the 2nd person singular. The further annotations, however, are function-oriented. Therefore, one would find the same item *babuska-ŋ* 'midwife-2SG' when searching for an agent (Semantic Roles), a subject (Syntactic Functions), a given referent (Information Status), a topic (Information Structure), or a cultural Russian borrowing (Borrowing). The function-oriented annotation tiers particularly contribute to the wide applicability of the corpus, since they enable the user to search specifically for these functional categories, even without having deep knowledge of the Dolgan language. This is relevant for typologists and/or theoretical linguists working with many languages and seeking for specific empirical data for their work. In order to illustrate this, Figure 2 shows the various annotations in a narrative text.

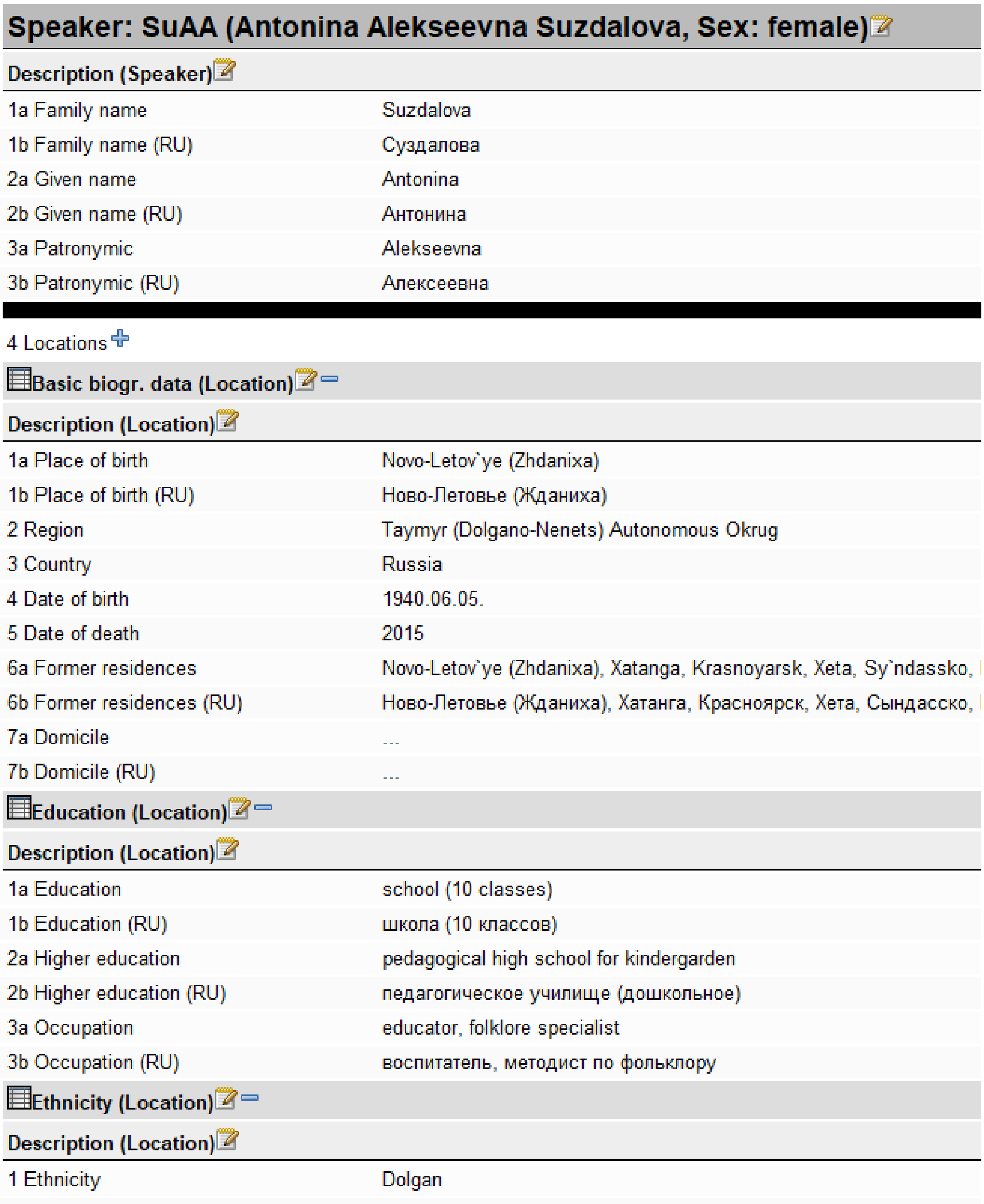
[13] The Leipzig Glossing Rules were developed and are regularly updated by the Max Planck Institute for Evolutionary Anthropology. The current version is available online at https://www.eva.mpg.de/lingua/resources/glossing-rules.php (last access: 27/04/2020).



	242 [02:17.0]	243 [02:17.5]	244 [02:	245 [02:18.6	246 [02:19.2]	247 [02:19.8]	248 [02:20.4]	249 [02:21.0]
ref	SuAA_20XX_Birth_nar.029 (001.029)							
st	Оччого буоллагына дуо бу баабускаҥ кэлэр ураhа дьиэгэ, төрүүр дьиэгэ, дьукаага.							
ts	Oččogo bu͡ollagina du͡o bu ba:buskaŋ keler uraha d’i͡ege, törü:r d’i͡ege, d’uka:ga.							
tx	Oččogo	bu͡ollagina	du͡o	bu	ba:buskaŋ	keler	uraha	d’i͡ege,
mb	oččogo	bu͡ollagina	du͡o	bu	ba:buska-ŋ	kel-er	uraha	d’i͡e-ge
mp	oččogo	bu͡ollagina	du͡o	bu	ba:biska-ŋ	kel-Ar	uraha	d’i͡e-GA
ge	then	though	MOD	this	midwife-2SG.[NOM]	come-PRS.[3SG]	pole.[NOM]	tent-DAT/LOC
gg	dann	aber	MOD	dieses	Hebamme-2SG.[NOM]	kommen-PRS.[3SG]	Stange.[NOM]	Zelt-DAT/LOC
gr	тогда	однако	MOD	этот	повитуха-2SG.[NOM]	приходить-PRS.[3SG]	шест.[NOM]	чум-DAT/LOC
mc	adv	ptcl	ptcl	dempro	n-n:(poss).[n:case]	v-v:tense.[v:pred.pn]	n.[n:case]	n-n:case
ps	adv	ptcl	ptcl	dempro	n	v	n	n
SeR	adv:Time				np.h:A			np:G
SyF					np.h:S	v:pred		
IST					giv-active			giv-active
Top					top.int.concr			
Foc						foc.int		
BOR					RUS:cult			
BOR-Phon					Vsub Csub			
BOR-Morph					dir:infl			
CS								

Figure 2: *Deep annotation in the INEL Dolgan Corpus*

The metadata of the corpus is stored in the *Corpus Manager (Coma)* component of the EXMARaLDA system. The metadata of transcripts (called “communications” in EXMARaLDA) contains information about the place and date of recording or the genre of the transcript, as well as information on who did what in the transcription, glossing and annotation. The metadata of speakers contains the basic biographical data of the relevant speaker, i.e., place and date of birth, education, language competence, ethnic composition of the family, place(s) of living, etc. Figure 3 shows an example of speaker metadata in the INEL Dolgan Corpus.



Speaker: SuAA (Antonina Alekseevna Suzdalova, Sex: female)

Description (Speaker)

1a Family name	Suzdalova
1b Family name (RU)	Суздалова
2a Given name	Antonina
2b Given name (RU)	Антонина
3a Patronymic	Alekseevna
3b Patronymic (RU)	Алексеевна

4 Locations

Basic biogr. data (Location)

Description (Location)

1a Place of birth	Novo-Letov`ye (Zhdanixa)
1b Place of birth (RU)	Ново-Летовье (Жданиха)
2 Region	Taymyr (Dolgano-Nenets) Autonomous Okrug
3 Country	Russia
4 Date of birth	1940.06.05.
5 Date of death	2015
6a Former residences	Novo-Letov`ye (Zhdanixa), Xatanga, Krasnoyarsk, Xeta, Sy`ndassko,
6b Former residences (RU)	Ново-Летовье (Жданиха), Хатанга, Красноярск, Хета, Сындасско,
7a Domicile	...
7b Domicile (RU)	...

Education (Location)

Description (Location)

1a Education	school (10 classes)
1b Education (RU)	школа (10 классов)
2a Higher education	pedagogical high school for kindergarden
2b Higher education (RU)	педагогическое училище (дошкольное)
3a Occupation	educator, folklore specialist
3b Occupation (RU)	воспитатель, методист по фольклору

Ethnicity (Location)

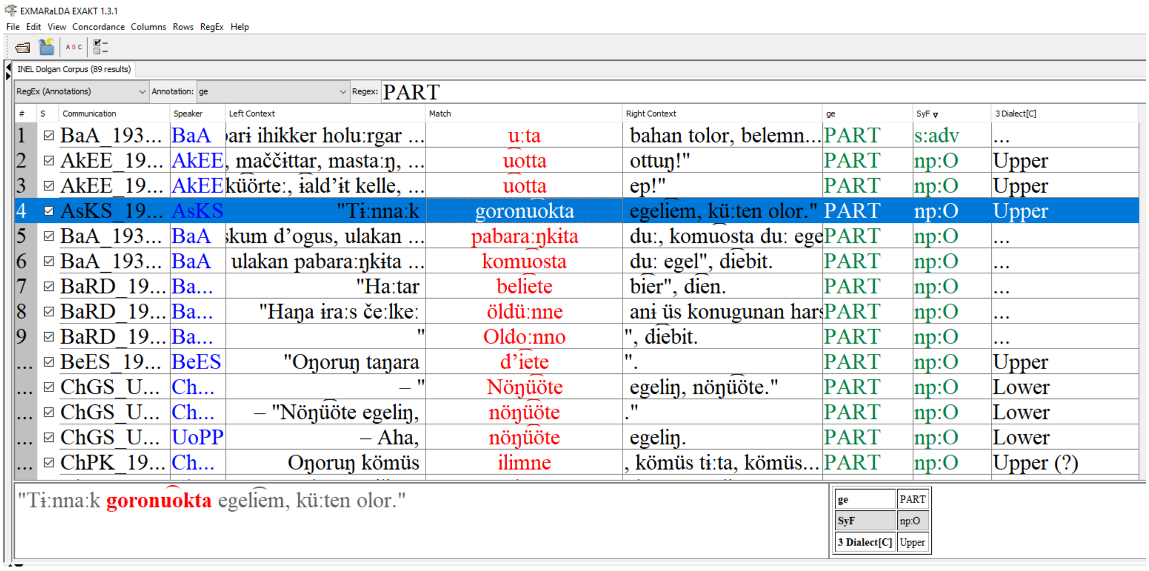
Description (Location)

1 Ethnicity	Dolgan

Figure 3: *Speaker metadata in the INEL Dolgan Corpus*

As was mentioned above, the INEL Dolgan Corpus can be searched using either the EXAKT tool form the EXMARaLDA program package or the web-based search via the Tsakonian Corpus Platform. Each tool has respective strengths. In EXAKT (Figure 4), concordance searches can easily be combined with metadata automatically extracted from COMA (see above). In Figure 4, a test-search for the partitive case in Dolgan is presented. As can be seen, the respective token (marked red) is shown within its context. Additionally, further columns with annotations and/or metadata can be included. Here, the annotation of syntactic functions (mostly NP objects) and the dialect of the given text

(Upper vs. Lower) was chosen. The concordance could be filtered for any value within these annotations, e.g., one could display only those tokens that come from the Upper Dolgan dialect.


EXMARaLDA EXAKT 1.3.1
File Edit View Concordance Columns Rows RegEx Help
INEL Dolgan Corpus (89 results)
RegEx (Annotations)
Annotation: ge
Regex: PART

#	S	Communication	Speaker	Left Context	Match	Right Context	ge	SyF	3 Dialect[C]
1	☑	BaA_193...	BaA	ɔarɨ ihikker holu:rgar ...	u:ta	bahan tolor, belemn...	PART	s:adv	...
2	☑	AkEE_19...	AkEE	, maččɨttar, masta:ŋ, ...	u͡otta	ottuŋ!"	PART	np:O	Upper
3	☑	AkEE_19...	AkEE	kü͡örte:, ɨ͡ald’ɨt kelle, ...	u͡otta	ep!"	PART	np:O	Upper
4	☑	AsKS_19...	AsKS	"Tɨ:nna:k	goronu͡okta	egeli͡em, kü:ten olor."	PART	np:O	Upper
5	☑	BaA_193...	BaA	ɔkum d’ogus, ulakan ...	pabara:ŋkɨta	du:, komu͡osta du: ege	PART	np:O	...
6	☑	BaA_193...	BaA	ulakan pabara:ŋkɨta ...	komu͡osta	du: egel", di͡ebit.	PART	np:O	...
7	☑	BaRD_19...	Ba...	"Ha:tar	beli͡ete	bi͡er", di͡en.	PART	np:O	...
8	☑	BaRD_19...	Ba...	"Haŋa ɨra:s če:lke:	öldü:nne	anɨ üs konugunan hars	PART	np:O	...
9	☑	BaRD_19...	Ba...	"	Oldo:nno	", di͡ebit.	PART	np:O	...
...	☑	BeES_19...	BeES	"Oŋoruŋ taŋara	d’i͡ete	".	PART	np:O	Upper
...	☑	ChGS_U...	Ch...	– "	Nöŋü͡öte	egeliŋ, nöŋü͡öte."	PART	np:O	Lower
...	☑	ChGS_U...	Ch...	– "Nöŋü͡öte egeliŋ,	nöŋü͡öte	."	PART	np:O	Lower
...	☑	ChGS_U...	UoPP	– Aha,	nöŋü͡öte	egeliŋ.	PART	np:O	Lower
...	☑	ChPK_19...	Ch...	Oŋoruŋ kömüs	ilimne	, kömüs tɨ:ta, kömüs...	PART	np:O	Upper (?)

"Tɨ:nna:k **goronu͡okta** egeli͡em, kü:ten olor."

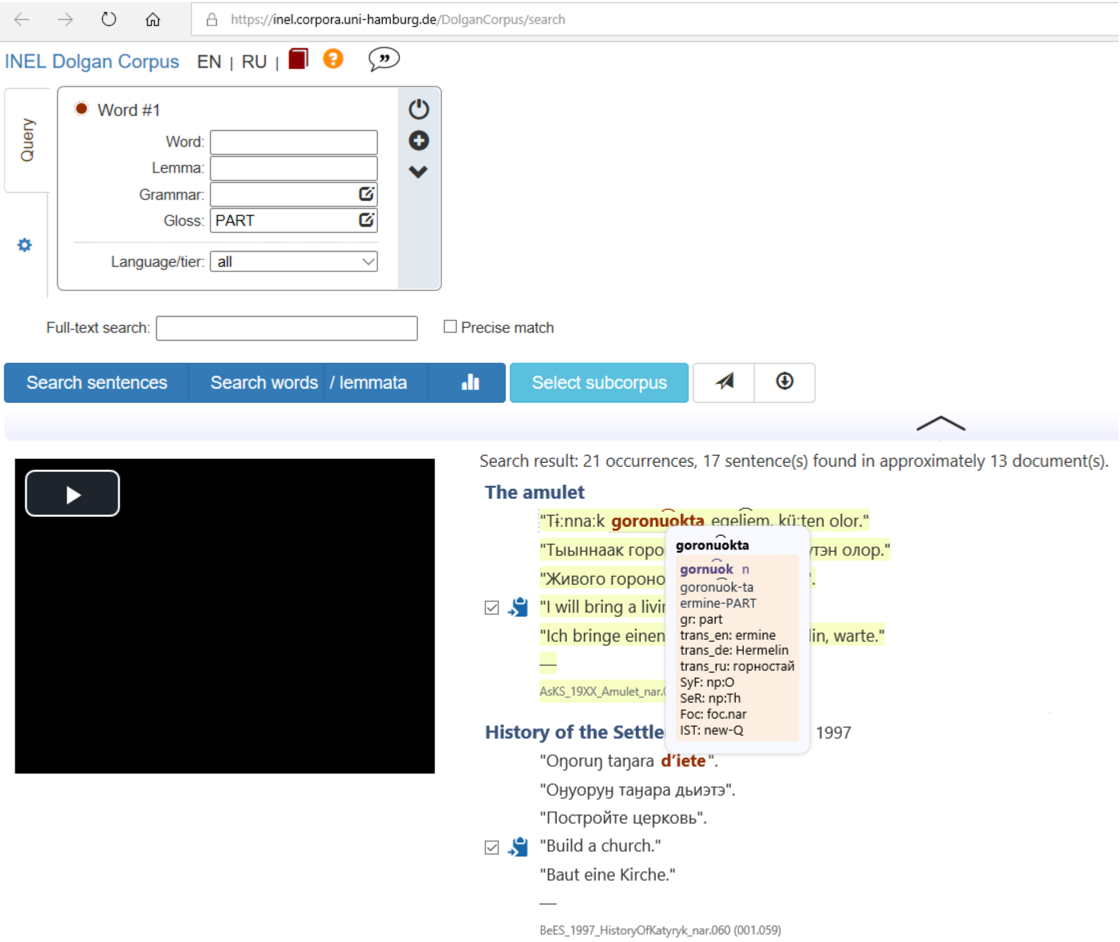
ge	PART
SyF	np:O
3 Dialect[C]	Upper

Figure 4: *Concordance search in EXAKT*

The Tsakonian Corpus Platform, in turn, has the advantage that it is web-based and does not require the whole corpus to be downloaded and stored locally. Additionally, it directly links the given token with its sound. By placing the cursor over the token, further information and annotations are given, if available in the respective transcript. In Figure 5 below, the same test-search for partitive singular is shown using the Tsakonian Corpus Platform.

Finally, it should be mentioned here that native speakers of Dolgan were involved in the work as much as possible, as it is the case for other languages, too. Here, it is especially noteworthy that Nina Kudryakova (the person responsible for Dolgan culture and folklore in TDNT), together with her relatives, transcribed and translated large parts of the TDNT material into Russian very reliably and quickly, using the intuitive and user-friendly software SayMore.[14] Without this collaboration, the amount of material included in the corpus would probably have been smaller. Additionally, Chris Lasse Däbritz and Eugénie Stapert (as a research fellow) conducted four weeks of fieldwork in Dudinka in summer 2017. Working up to eight hours with Dolgan informants per day, this fieldwork brought the project significantly forward, especially when it comes to clarifying uncertainties in texts and grammar; furthermore, they transcribed a great deal of material from Eugénie Stapert's collection.

[14] https://software.sil.org/saymore/, last access: 27/04/2020.



https://inel.corpora.uni-hamburg.de/DolganCorpus/search

INEL Dolgan Corpus EN | RU |

Query

● Word #1

Word:

Lemma:

Grammar:

Gloss: PART

Language/tier: all

Full-text search:

□ Precise match

Search result: 21 occurrences, 17 sentence(s) found in approximately 13 document(s).

The amulet

"Tɨ:nna:k **goronu͡okta** egeli͡em. kü:ten olor."

"Тыыннаак горо… …тэн олор."

"Живого гороно… …."

☑ "I will bring a livi…

"Ich bringe einen… …lin, warte."

—

AsKS_19XX_Amulet_nar.…

goronu͡okta

gornu͡ok n
goronu͡ok-ta
ermine-PART
gr: part
trans_en: ermine
trans_de: Hermelin
trans_ru: горностай
SyF: np:O
SeR: np:Th
Foc: foc.nar
IST: new-Q

History of the Settle… 1997

"Oŋoruŋ taŋara **d’iete**".

"Оҥуоруҥ таҥара дьиэтэ".

"Постройте церковь".

☑ "Build a church."

"Baut eine Kirche."

—

BeES_1997_HistoryOfKatyryk_nar.060 (001.059)

Figure 5: *Concordance search using the Tsakonian Corpus Platform*

4 Conclusion

The publication of the INEL Dolgan Corpus fills a considerable gap in the documentation and investigation of this under-studied language. It is now possible to conduct empirically based research on Dolgan, irrespective of the object of interest and/or the theoretical approach applied. Several studies (e.g., Däbritz 2018, Däbritz 2019) have already made use of this methodological advantage. We hope that the INEL Dolgan Corpus will encourage the linguistic community to conduct similar studies and to contribute as much as possible to the investigation of the Dolgan language.

Finally, the INEL Dolgan Corpus – as well as the other INEL corpora – may hopefully show that language documentation and corpus building projects do not necessarily depend on gathering new linguistic material. In many cases, especially when it comes to the indigenous languages of the Russian Federation, there is already very valuable material that "waits" to be located and worked upon – the INEL project may be a kick-off and an inspiration for projects having comparable agendas in the field of Uralic languages, and beyond.

References

Arkhangelskiy, Timofej, Anne Ferger & Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In: *Proceedings of the Fifth Workshop on Computational Linguistics for Uralic Languages*, 115–124. Available online: https://www.aclweb.org/anthology/W19-0310.pdf

Arkhipov, Alexandre. 2020. *INEL Corpora General Transcription and Annotation Guidelines*. In: *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 5. Szeged & Hamburg: University of Szeged, Department of Finno-Ugric Studies & Universität Hamburg, Zentrum für Sprachkorpora.

Arkhipov, Alexandre & Chris Lasse Däbritz. 2018. Hamburg Corpora for Indigenous Northern Eurasian Languages. *Tomsk Journal of Linguistics and Anthropology* 3 (21). 9–18.

Brykina, Maria, Svetlana Orlova & Beáta Wagner-Nagy. 2020. INEL Selkup Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmberg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 30/06/2020. Archived in Hamburger Zentrum für Sprachkorpora. http://hdl.handle.net/11022/0000-0007-E1D5-A

Däbritz, Chris Lasse. 2018. Predicative possession in Dolgan. *Tomsk Journal of Linguistics of Anthropology* 2 (20). 29–38.

Däbritz, Chris Lasse. 2019. First person imperative in Dolgan – Clusivity or number distinction? *Finnisch-Ugrische Mitteilungen* 43. 1–12.

Däbritz, Chris Lasse. 2020. *User's Guide to INEL Dolgan Corpus*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 4. Szeged & Hamburg: Department of Finno-Ugric Studies of the University of Szeged & Universität Hamburg, Zentrum für Sprachkorpora. https://doi.org/10.14232/wpcl.2020.4.

Däbritz, Chris Lasse, Nina Kudryakova & Eugénie Stapert. 2019. INEL Dolgan Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmberg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 31/08/2019. http://hdl.handle.net/11022/0000-0007-CAE7-1. Archived in Hamburger Zentrum für Sprachkorpora.

Efremov, Prokopij E. et al. (eds.) 2000. *Fol'klor Dolgan*. Pamyatniki fol'klora narodov Sibiri i Dal'nego Vostoka 19. Novosibirsk: Izdatel'stvo Instituta Arkheologii i Etnografii Sibirskogo Otdelenija Rossijskoj Akademii Nauk.

Gusev, Valentin, Tiina Klooster & Beáta Wagner-Nagy. 2019. INEL Kamas Corpus. Version 1.0. In: Wagner-Nagy, Beáta, Alexandre Arkhipov, Anne Ferger, Daniel Jettka & Timm Lehmberg (eds.). *The INEL corpora of indigenous Northern Eurasian languages*. Publication date 15/12/2019. http://hdl.handle.net/11022/0000-0007-DA6E-9. Archived in Hamburger Zentrum für Sprachkorpora.

Ubrjatova, Elizaveta I. 1985. *Jazyk Noril'skich Dolgan*. Novosibirsk: Nauka.

VPN 2010 = *Vserossijskaja perepis' naselenija 2010*. 4. Nacional'nyj sostav i vladenie jazykami [All-Russian census 2010. 4. National composition and command of languages]. http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf, last access: 22/04/2020.

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 1. Szeged & Hamburg: Department of Finno-Ugric Studies of the University of Szeged & Hamburger Zentrum für Sprachkorpora der Universität Hamburg. https://doi.org/10.14232/wpcl.2018.1

Chris Lasse Däbritz
Institute for Finno-Ugric/Uralic Studies, University of Hamburg
chris.lasse.daebritz@uni-hamburg.de