Kola Saami Christian Text Corpus

Michael Rießler
University of Eastern Finland
michael.riessler@uef.fi

Abstract

Christian texts have been known to be printed in Kola Saami languages since 1828; the most extensive publication is the Gospel of Matthew, different translations of which have been published three times since 1878, most recently in 2022. The Lord's Prayer was translated in several more versions in Kildin Saami and Skolt Saami, first in 1828. All of these texts seem to go back to translations from Russian. Such characteristics make these publications just right for parallel text alignment.

This paper describes ongoing work with building a Kola Saami Christian Text Corpus, including conceptional and technical decisions. Thus, it describes a resource, rather than a study. However, computational studies based on these data will hopefully take place in the near future, after the Kildin Saami subset of this corpus is finished and published by the end of 2024.

In addition to computation, this resource will also allow for comparative linguistic studies on diachronic and synchronic variation and change in Kola Saami languages, which are among the most endangered and least described Uralic languages.

1 Religious text production in Kola Saami languages

Religious texts constitute a significant part of the earliest documented data for all four Kola Saami languages, chiefly translations of Christian texts which started to be created in the same period of time for Akkala Saami, Kildin Saami, and Skolt Saami. (No similar Christian texts are known to exist for Ter Saami, though.) They include the Lord's Prayer and the complete Gospel of Matthew, each in different languages and versions, but also several other texts. The oldest text is from 1828, the youngest from 2022; the references of this paper include a full list of sources.

Notable are the recent texts created by Alexandra Antonova, in particular her Kildin Saami translation of Arapović's Jesus Friend of Children – a shorter version of Children's Bible including Lord's Prayer – and her completely new Kildin Saami translation of the Gospel. This text includes two different new translations of Lord's Prayer printed at the end of the book together with a translation of Apostles' Creed.

This book also includes a preface written in Kildin Saami by a non-Saami author. This text is relevant too because its language uses Christian metaphors and Christian symbols are explained, while using biblical terminology. Another relevant text is a prayer written by Saami author Jekaterina Korkina in Kildin Saami and Russian, with which she introduced a literary publication of hers (Korkina 2005).

These new texts not only add data to the corpus in terms of quantity, but allow for interesting comparative linguistic studies into various dimensions. This is particularly true because the idiolect of the recent translator Antonova (born 1932) is more than 100 years younger than that of Arvid Genetz's native speaker informant Parfenty Pyanov (born 1821), while both speakers seem to have the same dialectal background due to their family ties in the original Kiillt siida. Furthermore, the comparison between Antonova's two translations of different New Testament texts – created within a period of about a decade, but including interesting deviations in spelling and terminology – may potentially unravel some linguistic mysteries around her own derivation of the Kildin Saami orthography standard. Note also that the fragment of an intermediate manuscript version of Antonova's transla-

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 138–144
November 28-29, 2024 ©2024 Association for Computational Linguistics

Table 1: Currently included texts

Year		Text	Language	Word tokens	Status
1826	(1828)	Lord's Prayer	Kildin	59	Finished
1826	(1828)	Lord's Prayer	Skolt		Planned
1826	(1828)	Lord's Prayer	Skolt		Planned
1876	(1878)	Lord's Prayer	Kildin	60	Finished
1876	(1878)	Matthew (1–22)	Kildin	13,114	Not proofread
1876	(1878)	Matthew (23–28)	Akkala	5,014	Not proofread
1876	(1879)	Lord's Prayer	Kildin	62	Finished
1876	(1879)	Matthew (1–22)	Kildin	13,149	Not proofread
1876	(1879)	Matthew (23–28)	Akkala	5,001	Not proofread
≤1894	(1894)	Matthew (1–28)	Skolt		Planned
≤1895	(1895)	Lord's Prayer	Skolt		Planned
≤1895	(1895)	Primer	Skolt		Planned
≤1996	(1996)	Jesus Friend of	Kildin	8,180	Finnished
≤1996	(1996)	Lord's Prayer	Kildin	63	Finished
1999	(1999)	Orthodoxy	Skolt		Planned
≤2008	(2010)	Matthew (1)	Kildin	322	Finished
2005	(2005)	Prayer	Kildin	76	Finished
≤2009	(2022)	Lord's Prayer	Kildin	60	Finished
≤2009	(2022)	Matthew (1–28)	Kildin	18,215	Not proofread
≤2014	(2022)	Apostles' Creed	Kildin	71	Finished
≤2014	(2022)	Lord's Prayer	Kildin	58	Finished
2022	(2022)	Preface	Kildin	559	Finished

tion (including the complete first chapter) is also available to corpus studies because it has been published in a research paper (Jermolajeva 2010).[1]

Furthermore, the Kildin Saami subcorpus includes a small amount of data relevant to study Kildin Saami learners' language. In addition to the abovementioned preface, written by L2 learner Scheller, lines 16:22 through 16:28 of Gospel of Matthew were translated by Scheller (born 1977) and Elvira Galkina (born 1965).[2] The first is a language researcher and language activist. The latter has become known as poet and author of children's books and song lyrics in Kildin Saami language, although she describes herself as having full L1 speaker proficiency only in Russian.[3]

Since literacy for the Akkala Saami has never been established and no newer written language published, the six chapters from the Gospel are the only existing orthographic texts in Akkala Saami.

Regarding Skolt Saami, the corpus is incomplete. New relevant texts have been produced in contemporary Skolt Saami, but at the current state of this research I have not had the time and resources to identify and catalogue all existent Skolt Saami texts. The only exception is a small pictorial dictionary named The What, Why, and How of Orthodoxy (Kasala 1999). Since this work has a parallel version in Finnish, it is perfectly suited for the current project.

In addition to incomplete coverage of Skolt Saami, I've also not yet systematically searched for secondary or tertiary reprints of original texts to include them in the cor-

[1] Approx. 50 individual words from Antonova's manuscript are also listed in a paper by Bakula (2016, pp. 18–19) and could potentially be used for comparison.

[2] Both have published a relatively significant amount of other texts in Kildin Saami, which are not Christian but are available and could potentially be compared too. See, e.g. the Wikidata Query Service (https://query.wikidata.org/), where relevant metadata for titles with Galkina resp. Scheller as author or translator can be found easily.

[3] See the archived version of her professional CV at http://web.archive.org/web/20240404155219/https://www.masu.edu.ru/special/fip-saami/files/CVГалкина.pdf (2024-04-04).

pus. But plenty of them exist, first of all Genetz's own reprints of his transcripts of the Gospel in Genetz (1879b) and Genetz (1891). But also his Lord's Prayer was reprinted in Bergholtz (1894). Sjögren's Lord's Prayers were reprinted in Dalton (1870). Note also the more recent reprint of all these earlier versions in Németh (1991).

2 Corpus data

All mentioned Saami text sources from the 19th century are in the public domain. The same is true for versions in other languages which are all potentially useful for text parallelization but not inlcuded here.

Some of the Kola Saami texts were available in digital form earlier, others were digitized and proofread by me. Also Markus Juutinen (University of Oulu) – with whom I exchanged significant parts of these data – digitized and proofread texts for his abovementioned study. In addition, I worked together with Sergey Nikolaev (a Saami from Russia, today living in Oulu, Finland). Later I started uploading texts to Wikisource,[4] where proofreading and indexing has since been continued with the help of collaborators, who I don't know personally.

Parts of the 1878 edition of the Gospel– currently including chapters 3 through 10 – have been structured and made available as a corpus by the Lingvodoc project led by Julia Normanskaja (ILS RAS, Moscow). This corpus is structured in XML (at the levels of chapter, verse, word, and bound morpheme) and includes the original orthography, a Russian translation, tokenization of the original orthography and a translation of each token in contemporary Kildin Saami, and a morphological interlinearization with glosses.[5]

The user rights for Jesus Friend of Children were cleared by the Language Bank of Finland already in 1989, when the printed book was digitized in a project led by Pirkko Suihkonen (University of Helsinki).[6] The digital data was stored in pre-processed form – including text files with pre-Unicode encoding and including OCR errors – when I got in contact with the Language Bank in 2015. I was allowed to create a working copy of the repository for my own research. After having fixed the encoding and rebuilt Antonova's spelling with the help of a Pearl script and additional manual correction, and gave my improved version of the corpus back to the colleagues in Helsinki. But unfortunately, the Kildin Saami data has still not been published by the Language Bank.

The user rights for the new translation of the Gospel have yet to be cleared,[7] but the copyright laws of Finland and the European Union principally allow the use of such data as research material – including communicating it as part of research activities – even without a specific agreement with the copyright holders. This includes the typical processes for text and data mining of printed texts: digitizing as well as digital storing and processing.[8] It is also legal to publishing fragments as data illustrations for the purpose of teaching or in scientific publications, like in this paper.

However, more specifically defined user rights for Antonova's translation of the Gospel will hopefully lead to an open corpus publication in the future. Ideally, this can be done using the functional user interface Korp, for instance at GiellaLT in Tromsø, which promotes Open Science and with whom Scheller has been collaborating for several years.[9] But also the Korp platform at the Language Bank

[4] See, for instance, the index for Kildin Saami: https://wikisource.org/wiki/Category:Кӣллт_са̄мь_кӣлл (2024-10-11).

[5] See https://lingvodoc.ispras.ru/corpora_all (2024-10-11). The resource consists of one single file and does neither include metadata about its origin or any specification of a user license. But the sole originator seems to be Viktoria Bakula, professsor and specialist of Kola Saami languages at Murmansk Arctic University.

[6] The metadata in the repository, dated July 10. 1998, specify that The texts of the computer corpus of Kildin Sámi have been donated to the University of Helsinki by the Institute for Bible Translation (Stockholm, Sweden) to be used as research material. Reference to the corpus has to be made in papers in which it is used as a source.

[7] According to the publisher, i.e. the Stockholm branch of Institute for Bible Translation, copyright is held by the correctors of the text (researchers Elisabeth Scheller at the Arctic University of Tromsø and Elvira Galkina at the Arctic University of Murmansk) and the legal heir of the translator (Antonova's son Sergey Antonov from Lovozero).

[8] This refers to the exceptions in the EU Directive on Copyright in the Digital Single Market, which apply to text and data mining in academic research. National laws in EU countries follow the Directive. The name of the relevant Finnish law is (in Swedish) Upphovsrättslag, see https://www.finlex.fi/sv/laki/ajantasa/1961/19610404 (2024-11-01).

[9] See https://sanj.oahpa.no/about/ (2024-10-11).

Table 2: Parallel text fragments from Lord's Prayer in Kildin Saami; the two versions from 1876 origin from the same spoken recording, which was first transcribed and later represented in Cyrillic orthography; all versions but the first represent the one and the same dialect.

Speaker	Dialect	Text	
1828 (unknown, b. ≤1800)	Arsjogk	[…] Paſs låndſj tono namme.	[…] Amin.
1876 (Pyanov, b. 1821)	Kiillt	[…] ḁ$_{i}$nn pa$_{i}$zxuv tōn' nomm;	[…] Amin.
1876 (Pyanov, b. 1821)	Kiillt	[…] ань пазьхув тонэ нэм,	[…] Амин
1996 (Antonova, b. 1932)	Kiillt	[…] святэ ля̄ннч нэ̄мм То̄н;	[…] Зоа̄бэль
2014 (Antonova, b. 1932)	Kiillt	[…] Я пассьлувант нэ̄мм То̄н;	[…] Аминь.
2022 (Antonova, b. 1932)	Kiillt	[…] Анѣ пассьювв нэ̄мм То̄н;	[…] Аминь.

of Finland – where similar parallel corpora for other Uralic languages are already available[10] – would be a logical option.

With the exception of Schekoldin's primer, all texts easily allow for alignment to parallel versions. These versions exist in between the Kola Saami languages: a) Lord's Prayer in Skolt Saami (currently 2 versions) vs Kildin Saami (5 versions, plus one orthographic derivation), b) Gospel of Matthew in Skolt Saami vs Kildin Saami (chapter 1; one version in Skolt Saami and three versions in Kildin Saami), c) Gospel of Matthew in Skolt Saami vs Kildin Saami (chapters 2–23; one version in Skolt Saami and two versions in Kildin Saami), d) Gospel of Matthew in Skolt Saami vs Akkala Saami (chapters 24–28, one version each), and e) Gospel of Matthew in Kildin Saami vs Akkala Saami (chapters 24–28, one version each). But all of them can also easily be aligned with other language versions of the same texts, first of all to the Russian sources of the Saami translations.

Another dimension for parallel alignment results from the fact that the Akkala Saami and Kildin Saami translations of the Gospel published by Genetz were first documented in phonemic script (first published 1879) and later normalized by Genetz in Cyrillic orthography (first published 1878).

The overview in Table (1) lists the subparts of the corpus and the currrent state of their completion (year refers to the date of origin (≤ marks a terminus ante quem), the data of first publication is shown in parenthesis; word tokens may be due to corrections later).

It seems that the very existence of these parallel Christian texts has been known in general, but not in detail by all researchers in the field. For instance, a set of phonological studies by Bakula (2016) and Normanskaja (2016)[11] ignores the existence of Pyanov/Genetz's 1876 translation of the Gospel as a phonemically exact transcript and builds on the orthographic version instead. This is an omission which made the results significantly less useful.

Also the work with the new translation of the Gospel would likely have profited from a more complete overview of earlier texts. Deducing from the description of the translation and edition process in Scheller (2022) the two text correctors (Scheller and Galkina) were not aware of all different earlier versions of the Lord's Prayer, not even Antonova's own. And Scheller doesn't mention in the preface that Antonova's earlier translation of New Testament texts would potentially be related to her new translation of Gospel of Matthew. See, for instance the Sermon on the Mount, which Antonova translated in two different versions. This may be counterintuitive for readers, even if both versions are idiomatic Kildin Saami.[12]

3 Technical procedures and conventions

Building this corpus has been carried out for two decades already as part of the author's

[10] See https://clarino.uib.no/comedi/editor/lb-2020021121 (2024-10-11).

[11] These papers were reprinted with minor modifications as Normanskaja and Bakula (2022) without reference to the original work.

[12] The 1996 translation by Antonova was published by the Helsinki branch of Institute for Bible Translation, which specializes in the Uralic languages of Russia and supported by an editorial team. The 2022 translation, published by the Stockholm branch, lacked resources for thematic editorial checks. They had to rely on the competence of the text originators and could only support typography and typesetting (Brane Kalcevic, email 2024-10-08).

work with the Kola Saami Documentation Project (KSDP)[13] but did not aim at more than a convenient corpus of interesting data samples until very recently. Work on this corpus has also never been funded by means of a specific project grant.

At present, the corpus is stored and versioned in a private GitHub repository,[14] because parts of it are protected by copyright and can only be shared with research collaborators.

All original texts have either been digitally copied from other repositories or digitized by means of OCR by me before being modelled in XML. The data is encoded in UTF-8.

XLM markup follows the conventions of KSDP (cf. Blokland et al. 2015, pp. 12–14). There are other, more common formats available for modelling corpus data nowadays than XML (e.g. JSON). But XML has been the format of choice for KSDP because its data already includes a large amount of speech recordings and even video recordings, all of which are annotated and time-aligned in XML with the help of the tool ELAN.[15] Adding written corpus data in the same structure (even though time-alignment is not relevant for written data) makes cross-corpus searches very simple. On the other hand, the used XML structure is consistent and well documented and can therefore easily be converted in other formats if future users prefer to do so.

Since the original intention of this project was different from digitization projects run by archival institutions or libraries, original pagination is not modelled in these corpus data. Also, all non-textual graphical details on the original pages are ignored because this corpus is aimed to serve linguistic research.

All texts are first chunked at the chapter level (if they are longer than one chapter). This chunking resulted in separate files which can be called "corpus sessions" (and which are conceptually equal to corpus sessions consisting of one continuous speech recording, e.g. an interview or a procedural, in the case of multimedia corpus data for Kola Saami). In the case of the Gospel, each corpus session is chunked for verses, in order to keeping the original indexes for parallelization.

The different versions of Lord's Prayer are not printed in one consistent verse structure, but manual alignment is simple for this short text and done based on verses Matthew 9:6–13 throughout all versions. Thus, parallel locations in the different versions of the Gospel – including Lord's Prayer – are linked to each other by means of a pointer to chapters and verses.

Whereas the verses in the Gospel are relatively long and often include several sentences, other texts are chunked for sentences. This is how I typically also chunk my other written corpus sessions because sentences are conceptually equal to utterances in my spoken corpus data.

Textual structure at larger levels (headers, empty lines, paragraphs, etc.) is modelled by means of additional markup, added by me in the text if needed.

No further lexical, morphological, or syntactic tagging of the corpus has been carried out so far. Currently, I focus on the consistent and complete structuring of the Kildin Saami parts and complete proofreading of the Akkala Saami and Skolt Saami parts. But inspection and even systematic filtering of many morphosyntactic forms is already possible using RegEx and lists of bound and free grammatical markers.

4 Preliminary linguistic observations

The parallel data in Tables (2) and (3) – illustrating language use in worlds almost 200 years apart from each other – clearly show that the Kildin Saami language has not changed substantially since 1826.

These versions are relatively similar in terms of syntax, morphology, and lexicon. But there are also differences, some of them may indicate language change, others are due to different choices by the translators, or perhaps translation errors. For instance, Antonova's syntax is clearly more involved than the older translation. Perhaps this is because it tries to reproduce the underlying Russian constructions.

Antonova had been translating very produc-

[13] A description of the early stages of this project is found in Rießler and Wilbur (2007).

[14] https://github.com/langdoc/KSCTC/(2024-11-21)

[15] ELAN was originally created for building, annotating, and searching multimedia corpora, see https://archive.mpi.nl/tla/elan (2024-10-18).

Table 3: Parallel text of Matthew 1:1 published 2022 (originators Antonova/Scheller/Galkina), 2010 (Antonova), and 1878 (Pyanov/Genetz) – compared with a North Saami translation from 1998.

Speaker	Language	Text (Matthew 1:1)
1878 (Pyanov)	Kildin	Isus Xristos, Dḁvi̊d a$_i$lk’, Ābram a$_i$lk’ pūldɵγ sā$_i$n’.
2010 (Antonova)	Kildin	Авраам Альк Давид Альк Иисус Христос пуллдэгк.
2022 (Antonova)	Kildin	Й̄исус Христос, Авраам Альк, Давид Альк пуллдэгк.
1998	North	Dát lea Jesus Kristusa, Dávveda bártni ja Abrahama bártni, sohka.

tively since the beginning of her writing in the 1980s. Her work is clearly based on the intuition of a fully proficient and active L1 speaker. But it is not much informed by earlier literary work, not even her own work. This can be seen in her different variants of Lord’s Prayer (Table 2). Note, for instance the creative translation of “Amen” with a Saami discourse marker in her 1996 version. This seems to originate from a sudden inspiration but was revoked again later and instead the Russian form of this declaration is used.[16]

Interesting is also the order of the possessive pronoun. Antonova puts it after the head noun like in the Russian original, even though the constituent order in Saami is much stricter than in Russian and would normally not allow this (see Table 2).

Also the comparison of the different versions of the Gospel reveals interesting findings. Already the very first sentence (1:1) is recorded in three different versions, including a fragment of the unpublished manuscript by Antonova which was mentioned by Scheller (2022). This sentence describes Jesus Christ’s descent after David and Abraham, thus in chronological relation to the Babylonian captivity. Syntactically, this sentence consists only of a noun phrase in all three versions (see Table 3). But it can be observed in these examples that constituent order is different in the old translation compared to Antonova’s former version. Whereas Antonova uses a strict head-final order even for all intermediate constituents (which seems consistent with archaic Saami and reconstructed Uralic syntax), Pyanov/Genetz put only the lexical nouns in head-final position. The proper nouns in the intermediate noun phrases are head initial (this syntax looks closer to Russian). Interestingly, in Antonova’s second version, the order of constituents is scrambled in a completely new way, which does not follow the logical content of the original biblical genealogy. This change in the constituent order may be due to a translation error, because Jesus Christ descends from David’s lineage (after the exile), who in turn descends from Abraham’s lineage (before the exile).

Thus, already in its current form, the Kola Saami Christian Text Corpus allows interesting studies on diachronic and synchronic variation and change in Kildin Saami. The next step will be the complete inclusion of the mentioned Akkala Saami and Skolt Saami texts. The availability of this resource will hopefully prompt new qualitative and quantitative linguistic studies on these Uralic languages in the future.

References

Arapović, Borislav (1996). Iiisus – paarrne kaannʹc. Stockholm: Institute for Bible Translation.

Bakula, Viktorija B. (2016). “Vokalizm p’ervogo sloga v kil’dinskom dialekte saamskogo jazyka po dannym Jevang’elija ot Matfeja (1878)”. In: Uralo-altajskije issledovanija 3.22, pp. 13–33.

Bergholtz, Gustaf Fredrik, ed. (1894). The Lord’s Prayer. In the principal languages, dialects and versions of the World. Chicago.

Blokland, Rogier et al. (2015). “Language documentation meets language technology”. In: IWCLUL 2015. Ed. by Tommi A. Pirinen et al. Tromsø: The University Library of Tromsø, pp. 8–18.

[16] The spelling of this word not as Аминҍ – with the so-called half-palatalization sign – clearly indicates this. In Kildin Saami, нь marks the voiced palatal nasal /ɲ/ which doesn’t occur in this word.

Dalton, Hermann, ed. (1870). Das Gebet des Herrn in den Sprachen Russlands. Kaiserliche Akademie der Wissenschaften.

Genetz, Arvid (1879a). “Orosz-lapp nyelvmutatványok. Máté evangélioma és eredeti textusok”. In: Nyelvtudományi közlemények 15.1, pp. 74–152.

– (1879b). Orosz-lapp nyelvmutatványok (Máté evangélioma és eredeti textusok). Budapest: Magyar Tudományos Akadémia.

– (1891). Wörterbuch der Kola-Lappischen Dialekte nebst Sprachproben. Helsingfors: Finska Vetenskaps-Societeten.

Gospoda mij Iisusa Christa Pas’ Jevangelie Matveest (1894). Archangel’sk.

Jermolajeva, A. S. (2010). “Sv’aščennye teksty v p’er’evod’e na saamskij jazyk (istoričeskij obzor)”. In: Bogoslovije, istorija i praktika misij. Ed. by A. B. Jefimov and L. N. Ivanova. Moskva: Izdatel’stvo PSTGU, pp. 111–117.

Kasala, Kalevi (1999). Ortodokslažvuõđ mâi′d, mõõzz, mä′httceerkavteâđ ǩeârjjaž. Ortodoklaž Noõri Lett.

Korkina, Jekaterina N. (Mar. 2005). “Ji′mmel!” In: Saa′m. Vaalt k rr′j-laasst 2, p. 3.

Maahtvjest Pa′ss Jevan′gelje (2022). Stockholm: Institute for Bible Translation.

Mah’tveest Pas’-Jevangeli (1878). Helsinki: British and Foreign Bible Society.

Németh, Zsigmond (1991). 96 gleiche Texte in uralischen Sprachen (Vaterunser). Szombathely.

Normanskaja, Julia V. (2016). “Jevangelije ot Matfeja (1878) kak pamjatnik točnoj fiksacii arhaičeskogo sostojanija kil’dinskogo saamskogo jazyka”. In: Uralo-altajskije issledovanija 3.22, pp. 34–45.

Normanskaja, Julia V. and Viktorija B. Bakula (2022). “Jevangelije ot Matfeja na saamskom jazyke kak dokazat’el’stvo točnosti prasaamskoj r’ekonstrukcii J. Lehtiranta”. In: Kirilličeskije pamjatniki na ural’skih i altajskih jazykah. 1. Grafikofon’etičeskije osob’ennosti knig XIX v. Ed. by Julia V. Normanskaja. Moskva, pp. 18–38.

Ođđa Testamenta (1998). Oslo: Norgga Biibbalsearvi.

Rießler, Michael and Joshua Wilbur (2007). “Documenting the endangered Kola Saami languages”. In: Språk og språkforhold i Sápmi. Ed. by Tove Bull et al. 11. Berlin: Humboldt University of Berlin, pp. 39–82.

Ščekoldin, Konstantin (1895). Azbuka dlja loparej, živuščich v Kol’skom uezde Archangel’skoj gubernii. Archangel’sk.

Scheller, Elisabeth (2022). “Eevvtlessaa′nn”. In: Maahtvjest Pa′ss Jevan′gelje. Stockholm: Institute for Bible Translation, pp. 3–5.

Sjögren, Andreas Johan (1828). Anteckningar om församlingarne i Kemi-Lappmark. Helsingfors: J. Simelii Enka.