Kildin Saami-Russian-(English) Parallel Corpus Building

Evan Hansen
University of Eastern Finland
Joensuu, Finland
evan.hansen@uef.fi

Abstract

This paper presents two parallel corpora of written Kildin Saami and the process of their compilation. The first, a dictionary corpus, contains 101,889 Kildin Saami tokens of example phrases/sentences from three Russian-Kildin Saami dictionaries and the glossary of the nonfiction book *Saami ornaments*, accompanied by the examples' respective headwords and translations into up to four other languages. Headwords where possible are paired with their underived base, making it a suitable resource for investigating questions surrounding morphological derivation in Kildin Saami. The second corpus comprises 23,884 Kildin Saami tokens and was compiled from *Saami ornaments*, a trilingual (Russian-Kildin Saami-English) book introducing various Saami handicrafts and their creators from across Russian Sápmi.

1 Dictionary Corpus

This first corpus was originally built to facilitate the study of morphological derivation in Kildin Saami in my master's thesis on the subject of the reflexivizer -Эдт- (-*edt*-). A description of the corpus is found in Hansen (2025, 38-43). The work described here aimed at enriching the corpus and comprises the example phrases and sentences featured in three bilingual Kildin Saami-Russian dictionaries: Antonova (2014), Afanas'eva et al. (1985), and Kert (1986), as well as the trilingual Kildin Saami-Russian-English glossary entries from Mozolevskaja and Mechkina (2015).

In all, the corpus contains 21,800 Kildin Saami examples, composed of 101,889 tokens at roughly 4.7 tokens per example. Due to as of yet unnormalized orthographic variation across the source material, it is not yet possible to reliably estimate the number of types or unique headwords, nor can the number of headwords unique to each source be given.

The corpus is currently stored in a private GitHub repository[1].

Table 1: Dictionary corpus contents

Source	Examples
Afanas'eva et al. (1985)	10,160
Antonova (2014)	10,641
Kert (1986)	230
Mozolevskaja and Mechkina (2015)	849

1.1 Corpus structure

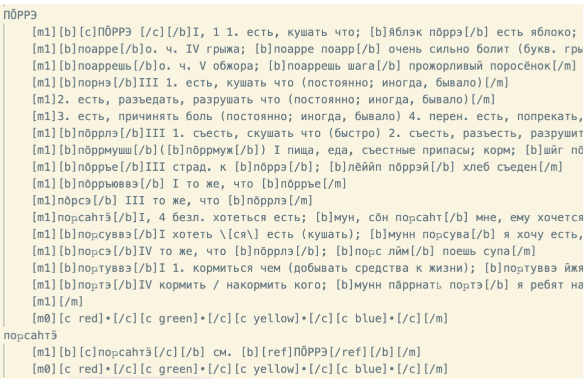
The corpus is in .csv format in UTF-8 encoding and contains 15 columns of data. The first five keep track of the source text, the associated headword, the headword in normalized orthography, the headword string in reverse, and the underived base headword of the aforementioned headword. Following these are two columns for the Kildin Saami (normalized orthography and original), then columns for Russian, Finnish, English, and German translations of the Kildin Saami example phrases. Lastly, there are columns for notes, inflectional information of the headword, and English and Russian definitions of the headword.

1.2 Data source selection

The three source dictionaries were chosen for a few reasons, the first being that they were accessible in .dsl format, a type of structured text file that can be used to compile dictionaries. The data being structured in this way made it straightforward to query and extract relevant data when corpus building, and it was additionally possible to load and visualize them simultaneously in the dictionary application Alpus. It was later a simple task to incorporate the data from Mozolevskaja and Mechk-

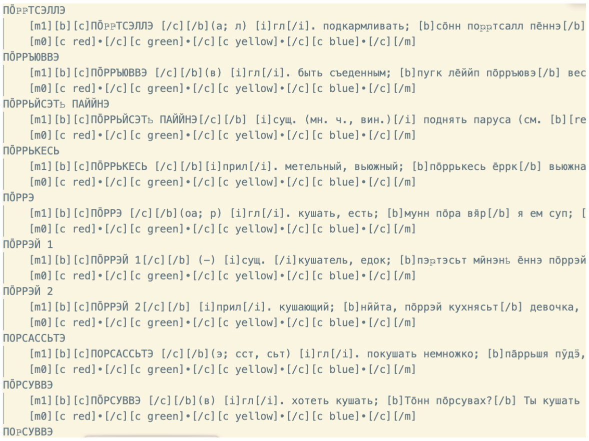
[1] Inquiries for accessing this resource should be directed to the data admins of the GitHub organization *langdoc*. The corpus is housed in the *sjd-parallel-corpus* repository (https://github.com/langdoc/sjd-parallel-corpus), which is currently set to "private" as it is a work in progress.

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 49–56
December 10-12, 2025 ©2025 Association for Computational Linguistics



```
ПŌРРЭ
    [m1][b][c]ПŌРРЭ [/c][/b]I, 1 1. есть, кушать что; [b]ӣблэк пōррэ[/b] есть яблоко;
    [m1][b]поарре[/b]о. ч. IV грыжа; [b]поарре поарр[/b] очень сильно болит (букв. гры
    [m1][b]поаррешь[/b]о. ч. V обжора; [b]поаррешь шага[/b] прожорливый поросёнок[/m]
    [m1][b]порнэ[/b]III 1. есть, кушать что (постоянно; иногда, бывало)[/m]
    [m1]2. есть, разъедать, разрушать что (постоянно; иногда, бывало)[/m]
    [m1]3. есть, причинять боль (постоянно; иногда, бывало) 4. перен. есть, попрекать,
    [m1][b]пōррлэ[/b]III 1. съесть, скушать что (быстро) 2. съесть, разъесть, разрушит
    [m1][b]пōррмушш[/b]([b]пōррмуж[/b]) I пища, еда, съестные припасы; корм; [b]шӣг пō
    [m1][b]пōррье[/b]III страд. к [b]пōррэ[/b]; [b]лēййп пōррэй[/b] хлеб съеден[/m]
    [m1][b]пōрръюввэ[/b] I то же, что [b]пōррье[/m]
    [m1]пōрсэ[/b] III то же, что [b]пōррлэ[/m]
    [m1]поҏсаhтӭ[/b]I, 4 безл. хотеться есть; [b]мун, сōн поҏсаht[/b] мне, ему хочется
    [m1][b]поҏсуввэ[/b]I хотеть \[ся\] есть (кушать); [b]мунн поҏсува[/b] я хочу есть,
    [m1][b]поҏсэ[/b]IV то же, что [b]пōррлэ[/b]; [b]поҏс лӣм[/b] поешь супа[/m]
    [m1][b]поҏтуввэ[/b]I 1. кормиться чем (добывать средства к жизни); [b]поҏтуввэ ӣжя
    [m1][b]поҏтэ[/b]IV кормить / накормить кого; [b]мунн пāррнатҍ поҏтэ[/b] я ребят на
    [m1][/m]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
поҏсаhтӭ
    [m1][b][c]поҏсаhтӭ[/c][/b] см. [b][ref]ПŌРРЭ[/ref][/b][/m]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
```

Figure 1: A screenshot from the .dsl file of Afanas'eva et al. (1985). The upper nest of entries contains пōррэ 'to eat' and its derivations. Below is the individual entry for порсаhтӭ 'to be hungry', which includes a reference back to пōррэ.



```
ПŌҎҎТСЭЛЛЭ
    [m1][b][c]ПŌҎҎТСЭЛЛЭ [/c][/b](а; л) [i]гл[/i]. подкармливать; [b]сōнн поҏҏтсалл пēннэ[/b]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЪЮВВЭ
    [m1][b][c]ПŌРРЪЮВВЭ [/c][/b](в) [i]гл[/i]. быть съеденным; [b]пугк лēййп пōрръювэ[/b] вес
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЬЙСЭТҌ ПАЙЙНЭ
    [m1][b][c]ПŌРРЬЙСЭТҌ ПАЙЙНЭ[/c][/b] [i]сущ. (мн. ч., вин.)[/i] поднять паруса (см. [b][re
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЬКЕСЬ
    [m1][b][c]ПŌРРЬКЕСЬ [/c][/b][i]прил[/i]. метельный, вьюжный; [b]пōррькесь ēррк[/b] вьюжна
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЭ
    [m1][b][c]ПŌРРЭ [/c][/b](оа; р) [i]гл[/i]. кушать, есть; [b]мунн пōра вӣр[/b] я ем суп; [
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЭЙ 1
    [m1][b][c]ПŌРРЭЙ 1[/c][/b] (–) [i]сущ. [/i]кушатель, едок; [b]пэҏтэсьт мӣнэнҍ ēннэ пōррэй
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРРЭЙ 2
    [m1][b][c]ПŌРРЭЙ 2[/c][/b] [i]прил[/i]. кушающий; [b]нӣйта, пōррэй кухнясьт[/b] девочка,
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПОРСАССЬТЭ
    [m1][b][c]ПОРСАССЬТЭ [/c][/b](э; сст, сьт) [i]гл[/i]. покушать немножко; [b]пāррьшя пӯдӭ,
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПŌРСУВВЭ
    [m1][b][c]ПŌРСУВВЭ [/c][/b](в) [i]гл[/i]. хотеть кушать; [b]Тōнн пōрсувах?[/b] Ты кушать
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПОҎСУВВЭ
```

Figure 2: A screenshot from the .dsl file of Antonova (2014), showing individual entries for пōррэ 'to eat' and others in alphabetical order. Compare with Figure (1), in which there is a noticable nested entry structure. Also important are POS tags (e.g. гл. (rus. 'verb') and information indicating morphophonological changes occurring during inflection (e.g. for пōррэ, ablaut through ō (oo) ⇒ oa (å) and consonant gradation with pp (rr) ⇒ p (r)).

ina (2015), following its inclusion in the corpus described in Section (2). This subset of data was added due to its wealth of example sentences and accompanying headwords.

In the case of Afanas'eva et al. (1985), another benefit is the nested structure of entries, where each headword houses daughter entries (if applicable) for headwords derived from it. Outside of the nests, daughter entries are also each afforded individual entries in alphabetical order, with a reference tag to the underived base form. Figure (1)

This greatly simplified the task of identifying derivational pairs, especially given that the regular morphophonological processes in Kildin Saami that manifest alongside derivational word formation can obscure quite drastically these links. This is because the orthography of words can be affected not just word-finally, which occurs for instance due to consonant gradation, but also at the beginning of a word where vowels often undergo the change of ablaut. We can consider the following base word and its derivations: ӣл (*eal*) 'life', ēльсуввэ (*jel'suvve*) 'to want to live', ӣллсэ (*iillse*) 'to get on (in life).' The result of such orthographic shifts renders typical alphabetic organization of dictionary entries rather unsuitable for Kildin Saami and for investigations into derivation. Afanas'eva et al. (1985) was further instrumental in identifying derivational pairs for data points from the other three sources, which do not possess a nested structure or reference tags for derived headword entries.

1.3 More about the data sources

In Afanas'eva et al. (1985), the headwords are based on everyday spoken language and cover the many facets of Saami society, culture, and day-to-day living. The dictionary comprises around 8,000 headwords, among which according to the foreword are a certain number of Russian loanwords (Afanas'eva et al., 1985, 9).

In the preface of Antonova (2014, 5), it is stated that the dictionary was compiled for a few reasons. The work was initially to be a simple word list to accompany the 2013 Kildin Saami translation of *Pippi Longstocking* entitled Тāрьенч Кукесьсуххк (*Taar'jenjč Kukessuhk*), translated by Aleksandra Antonova from the Russian version (Lindgren, 2014). There was a need for this word list due to the fact that a significant portion of the words featured in the translation were absent from the dictionaries available at the time. These dictionaries were further asserted to be not widely accessible. It was then decided that a new dictionary was necessary, not only for making the Kildin Saami translation of *Pippi Longstocking* more accessible, but also to provide the language community with a more comprehensive dictionary that includes practical vocabulary for everyday communication. The lexicography team additionally sought to incorporate terms that are highly relevant to Saami culture. The dictionary comprises around 8,000 headwords.

The later dictionary Antonova and Scheller (2021–) building upon Antonova (2014) features many spelling corrections, as well as general spelling convention normalizations based on Afanas'eva et al. (1985) (e.g. ли (*li*) ⇒ лӣ (*lii*) '[3SG] is' and ӑммьсе (*aamm'c'e*) ⇒ ӑммьсэ (*aamm'ce*) 'to yawn'). The dictionaries appear to share example sentences/phrases. In any case, data from this later work are not included in the corpus as no API or other tool are available to facilitate their extraction.

Kert (1986, 5) in its turn is a bidirectional bilingual Kildin Saami-Russian dictionary, composed of 4,000 headwords. It is described as being based on the vocabulary used in primary school textbooks, also including cultural terms relevant to Saami life. Russian loan words, with some exceptions, are excluded.

From Mozolevskaja and Mechkina (2015), 674 headwords and 849 examples were extracted. See Section (2) for more information about the text.

1.4 Dictionaries as data sources

When considering the data, the prescriptive nature of dictionaries should be kept in mind. The examples and phrases in the data were originally constructed by the authors (an exception being proverbs, though these do exhibit a certain degree of variation based on preliminary observations) and provide the reader with examples of how the words can be used in context. The source texts assert other information as well, such as how a given word declines. As such, questions of language variation and frequency are not feasible.

In addition to the above, a further consideration is that with the latest source publication (Mozolevskaja and Mechkina, 2015) being from 2015, there is roughly a decade of language change that is not featured.

1.5 Corpus compilation tools

The corpus building process largely took place in the Visual Studio Code (VSCode) application. Python scripts facilitated the work with the data, at times incorporating the libraries pandas, csv, deepl, and re. Some of VSCode's build-in functions like the find-and-replace tool also streamlined the process. The dictionary file reader Alpus was further utilized to cross-check the intended structure of the source dictionary entries as needed. Alpus was useful also in the way that it could query the dictionaries simultaneously, and certain orthographic discrepancies between the works would be ignored.

1.6 Data preprocessing

To extract the required data from the .dsl[2] dictionary files, it was mainly necessary to modify the organizational structures used to house the data. These were by-and-large strings that resemble HTML tags (for instance '[b]' and '[/b]') and were perhaps inserted by the original author for formatting the output. At times these disrupted the extraction process, sometimes cutting through data strings, or there were errors where tags were left open. A further complication was the fact that tags enveloping data also surrounded items like headwords or their inflected/declined forms. These difficulties aside, the tags also helped define boundaries for what needed to be extracted from the files.

Slight modifications to the code were needed as well due to some variations in structure between the dictionaries.

1.7 Data Extraction

When it came to the stage of extracting data, a list of headwords and their possible underived forms from Afanas'eva et al. (1985) was created. Using this list of pairs as a starting point, the examples from Antonova (2014) and Kert (1986) were extracted; as entries in these dictionary do not have references to their underived base words, it was possible to supplement these from pairs collected from Afanas'eva et al. (1985). It is however the case that the dictionaries use variations of the contemporary Cyrillic writing script employed for Kildin Saami, and it is possible that a few incorrect pairings were created across the dictionaries. Among the orthographic discrepancies are those relating to vowel length (e.g. т|оӑ|гктэдтэ (Antonova, 2014) ⇒ т|оа|гктэдтэ (Afanas'eva et al., 1985) and preaspiration (e.g. суэ|хх|птэдтэ (Antonova, 2014) ⇒ суэ|һ|птэдтэ (Afanas'eva et al., 1985).

These and other orthographic differences appear to be systematic for the most part but for the moment still have yet to be resolved.

Having established these headword-underived base pairs, they were matched up with their associated example phrases and sentences. The Kildin Saami examples were collected along with their Russian counterparts. The dictionary of origin was also recorded.

Later when revisiting the corpus for making im-

[2] These are freely available for download from https://slovari.saami.su/.

provements, all headwords *without* an associated example were added and paired with their underived base where possible.

1.8 Secondary translations

In an effort to make the data more accessible, machine translation was carried out on the data into English, Finnish, and German, as many publications related to Kildin Saami have been written in these languages. These translations were effectuated based on the Russian translations of the Kildin Saami entries given in the source dictionaries. The translations were made by means of the DeepL API Pro service and deepl Python library.

Though DeepL is one of the most robust machine translation services available, the quality and consistency of the translations appears to be not extremely reliable; this is regrettable, though it can be a helpful supplement to have access to when trying to quickly scan through the data.

As Kert (1986) and Mozolevskaja and Mechkina (2015) were not added to the corpus until later, they do not feature secondary translations.

1.9 Data postprocessing

At this stage, many errors were quite visible for the dictionary data in the way that for instance a tabulation error would generate an incorrect number of .csv columns across a block of data rows, and this could easily be detected both through even a brief glance and with the help of VSCode's built-in .csv error identifying features. Some errors also took the form of duplicate lines. The majority of large errors such as these were resolvable simply by using the find-and-replace tool in VSCode. Others required a bit more time and attention, where for example Russian data had somehow gotten into the Kildin Saami example column. Aside from these errors, some last escape characters and structural strings were removed.

One last postprocessing step concerned the secondary translations. With these certain translations that had for some reason been left blank by the API tool were redone. Another issue was related to grammatical and other tags in Afanas'eva et al. (1985) using abbreviated terminology. There are several dozen of these, such as страдательный залог 'passive voice' which appears as страд. in the dictionary entries. The Russian example data had been preprocessed by replacing the abbreviations with the corresponding expanded forms and enclosing them in asterisks, though this for many examples interfered with the translation done by the DeepL API. In total however, only around 500 of almost 21,000 examples contain one of these abbreviations, though not all translations have been addressed yet.

1.10 Future improvements

As mentioned in the above subsections, a notable issue with the corpus data is the lack of normalization of typos and of differences in orthography that can be found throughout the four source texts. This can perhaps first be fairly easily remedied for most cases of variation by targeting the *headwords* column, thereby making it possible to identify more underived roots for the headwords of those data points not originating from Afanas'eva et al. (1985). It would be most ideal however to eventually normalize the tokens in the examples as well, which would make it possible to explore derivations through corpus searches that include inflected forms.

A second improvement would largely affect the data originating from Antonova (2014). Among the headwords in this dictionary are entries for inflected forms of other headwords. These examples would be reassociated with the uninflected form.

A third key change for the future would be to assign at least one example sentence/phrase to those headwords in the data without one. These examples could in principle be copied from elsewhere in the data, as some headwords without examples do appear in those of others.

Next, the corpus will be continually enriched with new headwords as they are found. Kert (1988, 91) for instance includes the word рӯппсесьсиагаш (*ruupps'es'silagaš*) 'reddish' (from рӯппсесь (*ruupps'es'*) 'red'), which is not found in the corpus, nor is its adjectivizing suffix -сиагаш (*-silagaš*) '-ish' in combination with any other base. This is not the only attested derivational suffix missing from the corpus data, and it is crucial that such forms be represented for the corpus to be useful as a resource for exploring word-forming morphology in Kildin Saami.

Even within the corpus itself are words that are not given their own entries as headwords. The Russian loanword бутерброд (*but'erbrod*) 'sandwich' (same spelling and meaning in Russian) for example is used 10 times in the data but has no dedicated entry.

Finally, it should be mentioned that the data in the example sentences/phrases columns do not feature morphosyntactic tags. The data can already be queried fairly easily through RegEx searches, though inflected forms, in particular of underived words, remain a large hurdle. One possible avenue for data tagging would be to utilize GiellaLT's FST-analyser, although it is unclear whether this tool is available for research outside GiellaLT.[3]

2 *Saami ornaments* corpus

The second corpus at hand draws its contents from the 2015 publication entitled *Saami ornaments* (rus. Саамские узоры; sjd.[4] Сāмь кырьйнэз), a trilingual book that showcases Saami handicrafts from across 15 Saami siidas (sjd. сыййт, *syjjt*) within Russian Sápmi. In all, the corpus features 23,884 Kildin Saami tokens aligned with the Russian and English parallel texts. According to the front matter of the book, the original text is the Russian, which was subsequently translated into the Kildin Saami and English (Mozolevskaja and Mechkina, 2015, 2). The corpus is currently stored in a private GitHub repository (see [1]).

A significant portion of the data comes from the book's glossary, from which a total of 674 headwords and 849 examples were extracted. Of the examples, 55 are proverbs. What is more, entries appear to follow the inflection categorization system used in Afanas'eva et al. (1985), also borrowing certain example phrases. The glossary does however diverge at times from the aforementioned dictionary in orthography (in particular when it comes to long and short vowels), as well as with the inclusion of certain terms not found within Afanas'eva et al. (1985) or in the other two dictionaries referenced in Section (1).

2.1 Data structure

The corpus comprises 3,640 rows of data in .csv format, in UTF-8 encoding. There are two columns for the Kildin Saami data, one for the original data modified and the other for normalized orthography. Aligned with these are columns for the Russian and English parallel textual data. Further columns include one for notes, which mainly records the notes left in the margins of the book by those involved in the revising and editing process, as well as two columns for the sections and subsections to which the data belong. The sections reflect the overarching sections of the book, for instance the various Saami siidas and the "About the authors" portion. Subsections largely follow the headers for the many handicraft items in the book. Duplicate subsections within the same section are numbered. The final column is for I.D. numbers, which were added in order to keep track of the original internal structuring of the book.

2.2 Corpus building

Thanks to the availability of an OCR'ed copy of the book, as well as the book's highly structured contents with overall consistent use of (sub)headers and bulleted lists, the process of compiling this corpus was relatively straightforward.

The raw data first were copied and pasted into a .txt file section by section, keeping parallel language sections together. The data were then preprocessed using a Python script in order to have one sentence of data per line in the .txt file. Frequent problem strings such as '1.' and abbreviations like 'мн.' (rus. 'plural') were accounted for in the script to simplify the process. The data were then aligned in VSCode using the *Edit CSV* extension created by *janisdd*. The extension provides an Excel-like interface for reading and writing .csv files, allowing for simple data manipulation.

Concerning the alignment of the three versions, sentential alignment was prioritized. Instances where multiple sentences in one language mapped to one sentence in another were aligned as one data point, avoiding any divisions of single sentences. Rarely, and as necessary, sentences were rearranged for alignment.

Image captions and "Master of Sami Handicraft" boxes (sections providing the name of the craftsperson) were excluded for the most part, given their repetitive structure and redundant information.

2.3 Postprocessing

Once all data had been compiled, the data were checked for lookalike character replacements (e.g. the Latin <ä> (U+00E4) for the Cyrillic <ӓ> (U+04D3)) and other nonstandard characters. The biggest modification replaced the combining overlines (U+0305; e.g. <я̅>) used in the book to indicate a long vowel, this in place of the standard orthography's combining macron (U+0304; e.g. <я̄>).

[3] The imprint of GiellaLT's digital Kildin Saami dictionary mentions the existence of an automaton for paradigm generation, see `https://sanj.oahpa.no/about/`. But the GiellaLT infrastructure offers only an embryonic version, see `https://github.com/giellalt/lang-sjd`.

[4] Kildin Saami.

2.4 Preliminary observations

A notable aspect within the glossary is the inclusion of Varzino (Arsjogg) dialectal varieties for certain headwords. Some of these are given in Table (2). Note that the “standard”[5] variants are taken from Antonova (2014).

A further point of interest within the glossary are certain terms which are not listed in the three dictionaries included in the dictionary corpus outlined in Section (1). Among these are Е̄кесь Та̄ссьт (*iekes’ taass’t*) ‘Mars’, га̄рэс *(gaares)* ‘worsted yarn’, and быдтъесь *(bydtjes’)* ‘necessary’.

2.5 Future improvements

As of yet, the text in the normalized orthography data column has not undergone any normalization for orthographic variation or for possible typos, though this certainly is planned for the near future to allow for more streamlined corpus queries across multiple corpora; in this same vein, the corpus will soon be converted to XML format, as this has been preferred by the Kola Saami Documentation Project for such materials [6].

Table 2: Comparison of Standard and Arsjogg variants (NOM.SG)

Standard	Varzino (Arsjogg)	Translation
нЭ̄дт	нЭ̄ввт	’handle’
ка̄ӆӆт	ка̄ӆт	’pocket’
кыффкэмкарь	кыјјм-ка̄рръ	’mirror’
(or: кыйхмкарь)	”	
коалль	коалльт	’gold’

Certain segments of the data need to be revisited for alignment modifications. Though overall it was possible to identify correspondences across the three language versions, certain sentences appear to have been skipped over during the translation process. To some extent these unclear sections may be due to the corpus compiler’s competencies in Russian and Kildin Saami, which include structural knowledge and a novice L2 proficiency level in both languages.

3 Corpora applications

Turning first to the dictionary corpus, the structure makes it ideal for investigating questions related to derivation in Kildin Saami.[7] As derivation in the language is almost exclusively accomplished through suffixation, a key strength of the corpus is its column for headwords spelled in reverse order. Querying the data in a spreadsheet viewer, the rows of data can be sorted with this column to quickly identify all instances of a given suffix when it is the final derivation. This reversed spelling column being based on the normalized headword column further makes it possible to gather headwords with suffixes that vary orthographically by source (e.g. -аһтӭ and -аххьтэ; -нэх(х)ьк and -нэһкь).

Additionally, the ‘root’ column has been populated for many data points from Antonova (2014); Mozolevskaja and Mechkina (2015); Kert (1986) using derivational pairs created from Afanas’eva et al. (1985). This establishment of derivational relations for these data from other sources then saves the corpus user the time of having to identify the connections manually, not only for the parent-descendent pairings but also for the links between sibling headwords with a shared root word. The original nested headword structure from Afanas’eva et al. (1985) in effect is broadened, opening up opportunities to examine entire families of headwords side-by-side and pulling from multiple sources at once. This is particularly useful when comparing a headword carrying two or more layers of derivational suffixation with its possible intermediate headwords (e.g. шЭ̄ннтэ (*šeennte*) ‘to grow’ ⇒ шЭ̄ннтлэ (*šeenntle*) ‘to grow a little, grow quickly’ ⇒ шЭ̄ннтлуввэ *šeenntluvve*) ‘to start to grow, get taller’) or when considering aspectual derivations (e.g. шЭ̄ннтлэ (*šeenntle*), шЭ̄ннтассьтэ (*šeenntass’te*), шЭ̄нтсэ (*šeentse*) ‘to grow quickly’; шэнтнэ (*šentne*) ‘to grow continuously’).

Aside from derivation, patterns in Kildin Saami-Russian translation could be explored using the Russian columns. To investigate translation strategies involving desiderativity for instance, RegEx could be employed to return all rows of data with the Russian verb хотеть ‘to want’ and its variations. Important would be to also include the column containing Russian definitions of the Kildin Saami headwords, as not all have an accompanying Kildin Saami-Russian example phrase pair.

Through the incorporation of new headwords over time and more thorough orthographic normal-

[5] “Standard” here refers to the language as it is used in Afanas’eva et al. (1985), Antonova (2014), Kert (1986),and Mozolevskaja and Mechkina (2015). It is not explicitly stated which varieties are used in these texts.

[6] A description of the early stages of this project is found in Rießler and Wilbur (2007); see also Rießler (2024).

[7] My many thanks to the two anonymous reviewers who suggested including this section.

izations, the corpus will become a reference tool of increasing usage potential for those who develop community-facing resources, from spell-checkers to pedagogical materials. Lexicographers especially may benefit from the spelling normalization columns when deciding which forms to include in a dictionary and possibly even list multiple variations for users' ease of access and reduced prescriptivism. We can take as an example how Antonova (2014) features ōннъюввэ (*oonnjuvve*) while Mozolevskaja and Mechkina (2015) additionally uses оаннъюввэ (*ånnjuvve*), both meaning 'to be used' and derived from the verb оаннӭ (ånn'e) 'to use' though using a different stem.

As for the *Saami ornaments* corpus, the source material was selected for compilation with the intention of increasing the amount of multilingual parallel corpus data available for Kildin Saami. Particular to the source publication is its trilingual parallel structure, which makes it quite accessible to English-speaking researchers who have little to no proficiency in Russian and/or Kildin Saami. The corpus with its three parallel versions can serve as a starting point for questions pertaining to translation. Relatedly, quite many Russian loanwords are present in the data, some of which are not found in the dictionary corpus, like этнографическэ (*etnografičeske*) 'ethnographic'; focused studies on Russian borrowings may benefit from incorporating these data points into their research.

What is more, the data are valuable for their representation of contemporary nonfiction written language and their subject area of Saami handicrafts. With these attributes in mind, directions of research using the corpus could include analyses of vocabulary in relevant semantic domains (e.g. colors, materials, crafting tools/techniques) and language change through comparisons with source materials from other time periods. Researchers from the fields of literary studies, history, and anthropology may also have interest in the data.

4 A note on copyright

The two corpora described in this paper derive their contents from source texts that are protected under copyright. In principle, copyright laws within the European Union permit the use of the source materials for use in academic research; this includes converting them to a digital format, storing them, and processing them as is typically done when mining textual data.[8] It is furthermore permissible to conduct research with such materials in collaboration with other researchers.

Relatedly, fragments of the source material may be published in order to illustrate the data in teaching and scientific publication contexts. However, data extracted from material protected under copyright may not be made freely available, and for this reason the corpora are stored in a private repository.

Ideally, the two corpora would eventually be made more freely accessible, which could be accomplished with permission from the legal owners of the data.

References

Nina E. Afanas'eva, Aleksandra A. Antonova, Boris A. Gluchov, Lazar' D. Jakovlev, and Ekaterina I. Mečkina. 1985. *Saamsko-russkij slovar' = [Sām'-rūšš soagknehk'] = [Saami-Russian dictionary].* Russkij jazyk.

Aleksandra A. Antonova. 2014. *Saamsko-russkij slovar' =[Saami-Russian dictionary].* ANO Arktičeskij centr naučnich issledovanij i ėkspertiz.

Aleksandra A. Antonova and Elisabeth Scheller. 2021–. *Saamsko-russkij i Russko-saamskij slovar' =[Saami-Russian and Russian-Saami dictionary].* UiT The Arctic University of Norway.

Evan Hansen. 2025. Kildin Saami *-edt-* Reflexivized Verbs. Master's thesis, University of Eastern Finland.

Georgij Martynovič Kert. 1988. Slovoobrazovanie imen v saamskom jazyke =[Word formation of nouns in the Saami language]. In Georgij Martynovič Kert, editor, *Pribaltijsko-finskoe jazykoznanie. Voprosy leksikologii i grammatiki*, Trudy Karel'skogo Filiala Akademii Nauk SSSR, pages 84–91. Karel'skij filial AN SSSR.

Georgij Martynovič Kert. 1986. *Slovar' saamsko-russkij i russko-saamskij =[Saami-Russian and Russian-Saami dictionary].* Prosveščenie.

Astrid Lindgren. 2014. *Taar jenjč Kukessuhk [Pippi Longstocking].*

Anastasija E. Mozolevskaja and Ekaterina I. Mechkina. 2015. *Saamskie uzory =[Saam' kyr'jnez] =[Sami ornaments].* Drozdov-na-Murmane.

[8] The EU Directive 2019/790 on Copyright in the Digital Single Market outlines copyright exceptions that allow for text and data mining for scientific researcher purposes. This Directive is reflected in national laws within the EU.

Michael Rießler. 2024. Kola Saami Christian Text Corpus. In Mika Hämäläinen, Flammie Pirinen, Melany Macias, and Mario Crespo Avila, editors, *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 138–144. ACL.

Michael Rießler and Joshua Wilbur. 2007. Documenting the endangered Kola Saami languages. In *Språk og språkforhold i Sápmi*, pages 39–82.