In: Rudolf Muhr, Josep Angel Mas Castells, Jack Rueter (eds.) (2019): European Pluricentric Languages in Contact and Conflict: Wien et. al., Peter Lang Verlag. p. 201-208.

Jack RUETER / Mika Hämäläinen
(University of Helsinki, Finland)
jack.rueter@helsinki.fi
mika.hamalainen@helsinki.fi

Skolt Sami, the makings of a pluricentric language, where does it stand?

Abstract

This paper will provide a brief description of Skolt Sami and how it might be construed as a pluricentric language. Historical factors are identified that might contribute to a pluricentric identity: geographic location and political history; shortages of language documentation, and the establishment of a normative body for the development of a standard language.

Skolt Sami is assessed in the context of Sami languages and is forwarded as one of a closely related yet distinct language group.

Here the issue then becomes one of facilitating diversity even for under-documented languages. And we aptly describe opportunities in language technology that have been utilized to this end.

Finally, brief insight is given for other Uralic languages with regard to pluricentric character and possibilities for language users to facilitate the maintenance of their individual language needs.

1. Skolt Sami – A brief background

The Skolt Sami, a minority within a minority, speak a Sami language, which is a subbranch of the Uralic language family. Their geographic position places them on or near the borders of Finland, Norway and Russia.

The very emergence of Finland as a new nation in 1917 and the Peace of Tartu 1920 broke a centuries long tradition of an annual cycle. The Skolt Sami population was split between Russia and Finland with those living in Finland being forced to settle in the western reaches of their former annual cycle as migrant reindeer herders and fishers. In the aftermath of World War II, the population was once again splintered with the westward movement of borders. The change in the political affiliation from Russian to Finnish has seen a diminished Russian influence on the language and a marked attraction to the new majority language, Finnish.

The Skolt Sami language of today stems from two main geographic divisions (dialects): Paatsjoki and Suõ′nn’jel. Paatsjoki is further divided into the variants of Njauddâm (no. Neiden, extinct), Paaččjokk (fi. Paatsjoki), Peäccam (fi. Petsamo, ru. Печенге), and Mue′tǩǩ (fi. Muotka, ru. Мотко), while Suõ′nn’jel is further divided into the variants previously spoken in Suõ′nn’jel (fi. Suonikylä, ru. Приречный), Njuõ′ttjäu′rr (fi. Nuorttijärvi, ru. Нотозеро) and Sââ′rvesjäu′rr (fi. Hirvasjärvi, ru. Гирвасозеро).

Today, there are two large heterogeneous settlements with an aging population, whereas many of the active younger-generation language users and learners live and work in the multilingual center, Inari. All of these factors are makings for pluricentric language domains.

Skolt Sami is a distinct Sami language form (cf. Miestamo 2011; Fiest 2015). It is readily distinguished from North Sami, the largest Sami language both morphologically and lexically. Even though Skolt Sami might be distinguished from other Sami languages by the mere presence of special letters and glyphs not found in other Sami languages (ǩ, ǥ, ǧ, ′, ’, ʒ, ǯ, ä, å, õ), productive stem alternation is often a decisive element in paradigm cell distinction – a phenomenon reminiscent of the vowel variation found in the English pattern sing, sang, sung, song but more extensive and prominent. There is also a notable presence of Karelian and Russian loanwords, which may be attributed to a stronger Russian ecclesiastical influence from the past.

There are language forms deriving from three different countries, Finland, Norway, Russia. First, there is no surviving spoken form in or from Norway, though there are attempts being made to revive the Skolt Sami language and culture there (cf. Magga-Kumpulainen 2017). Second, the speakers of Skolt Sami in Russia do not have the same level of public infrastructure to support work in the language, although there is definitely interest, and there is one Skolt Sami speaker from Russia actively participating in the norming of the Skolt Sami language at Ǩiõlljuâggtõs (Skolt Sami Language division) in Inari, Finland at the Sami Parliament. Finland is the only country with ongoing publication activities in Skolt Sami.

The literary language norm is still being formulated. While the literary language norm has been set to correspond to the language form once spoken in Suõ′nn’jel, much of the traditional fieldwork documentation of the language has, in fact, given Paaččjokk a more prominent position. In the publication of dictionaries and wordlists, language form usage has inadvertently vacillated between the selected norm (Suõʹnnʼjel) and the language of some experts (Paaččjokk, Peäccam). A consensus is growing for maximal use of lexica from both major dialects with adherence to morphophonological characteristics from Suõʹnnʼjel (p.c. Merja Fofonoff, native language translator and interpreter).

Official language development as introduces translation and normative bodies. Translation work and language development is carried out by native speakers and language learners alike. At times, the conceptual space addressed in translation is quite removed from the traditional Skolt Sami domains, which results in the coining of new words alien to native speakers. The language forms spoken in the settlements is at odds with the language spoken in multilingual Inari.

Temporal and geographic distance from the speaking community has also seen a shift in terminology orientation. Where the older generation might be familiar with the word form biologii (a loan from Russian), the book makers might use the form biologia (a loan from Finnish). The direction of the language orientation is clear, a shift of orientation is underway in a language more and more integrated into the Finnish system.

Skolt Sami is difficult to see as a pluricentric language. It cannot be compared with Northern Sami, Lule Sami and Southern Sami, which have official status in more than one country. Whereas Northern Sami has official status in Norway, Sweden and Finland, and therefore must address the conceptual needs of three different forms of society on each side of the border, there is a consensus to consolidate language resources, which is exemplified in the existence of one mutual Sami-language television program for Norway, Sweden and Finland. There have, however, been inquiries made into the possibility of defining a separate Common locale data repository (cldr) for Northern Sami in Finland (pc. Finnish localization meeting, Helsinki, 26.9.2018).

To the outsider who is only aware of Sami as an intangible entity, learning there are nine living Sami languages with six established orthographies might be completely unexpected. One anecdote about the situation from the early 1980s was: “The more languages, the more professorships” (p.c. Raili Pirinen, Helsinki City Sámit). How is it possible that twenty to thirty thousand speakers of all nine language forms have their diverse vernaculars set off as separate languages? There is no simple answer. Although some of the languages might be seen as parts of three separate continuums, there are always questions of cultural distinctions as well, a good reason for scientific research and debate. More important, however, is the modern facilitation of this complex system, and how solutions can be applied for the revitalization of the smallest of the language groups.

2. Rule-based language technology

When only minimal lexical and text resources are available for a language, there is little possibility for statistical descriptions of a language. In fact, many languages explicitly studied as statistical phenomena are languages with minimal inflection. Where English has perhaps a maximum of 4 word forms per stem for only some parts of speech (nouns and verbs) and Mandrin Chinese even less, Skolt Sami has easily over two hundred forms per noun (diminutives, number, case, possessive suffixes, discourse particles), and this would implicitly mean that any statistically based research of an undocumented English corpus of 4 million words could only be equated to a Skolt Sami corpus of 200 million words. Since Skolt Sami does not have such research luxuries, we use rule-based description before applying anything reminiscent of statistical research or the even newer neural networks.

Rule-based language technology is one of the major features of the Giella (Northern Sami for Language) infrastructure centered at the Norwegian Arctic University in Tromsø, Norway, where language technological research (Giellatekno) and tool development (Divvun) meet.

It is in the Giella infrastructure that the lesser documented Skolt Sami language, accruing fieldwork outcomes and research results has been able to develop. Some of the most prominent materials can be found in the lexical and fieldwork of T.I. Itkonen (1958) and a the rapidly integrating work of Sammallahti and Moshnikoff (1991) as well as subsequent lexical documentation by Jouni and Satu Moshnikoff, being realized (2019-2020) in a collaborative project between the Saami Parliament in Inari and Giellagas-Instituutti, Oulu, Finland. Giella has also been a place where, more recently, a new dictionary system has been developed for allowing fuller application of morphological knowledge and lexical resources for the language user and professional, alike.

There are dictionaries based on the same materials but with diverse presentations and uses. Whereas the morphological dictionaries at (http://saan.oahpa.no/) allow for multiple language translation of individual word forms, there is also a click-in-text app that can be used on texts in browsers such as Firefox and Google Chrome, i.e. that means one can also use texts on the individual's own computer (see also Trosterud 2017). Additionally, these dictionaries provide direct search links to open-source research corpora on the (http://gtweb.uit.no/korp/) server, where Skolt Sami written materials categorized by genre, can be queried for advanced language learning by both the research and language community. Linking to other resource, e.g. the Indigenous language archives at the Giellagas-Instituutti are forthcoming.

The interconnected multilingual dictionary found through (https://mikakalevi.com/sanat/) and (https://www.akusanat.com/) provides the user external links to Álgu, for instance, an etymological data base of the Sami languages, as well as audio media archive materials at Max-Planck-Institut in Nijmegen, the Netherlands. The interconnected multilingual dictionaries also provide possibilities for crowd-sourcing (Rueter & Hämäläinen 2017), (Hämäläinen & Rueter 2018), sharing of conceptual space between dictionaries (Hämäläinen et al. 2018), and output for language technological resources for minority language facilitation.

Skolt Sami language technology is a part of the Giella infrastructure introduced in the Nodalida artcile “Building an open-source development infrastructure for language technology projects” (Moshagen et al 2013). The strong affiliation of Skolt Sami language technology with the other Sami languages provides parallels for reusable development of tools and language resources.

On the one hand, there are language tools, such as Intelligent Computer-Assisted Language Learning (ICALL: http://oahpa.no/nuorti/) as well as language-specific tools for multiple languages, e.g. keyboards, spellcheckers at (http://divvun.no/), and corpora derived from public domain translations of legal texts for the over the past two decades (http://gtweb.uit.no/korp/).

Language technology, however, is not a primary. It follows the lead of and collaborates with a normative body, which, in the case of Skolt Sami, is Ǩiõlljuâggtõs [Language Division] of the Sami Parliament in Inari, Finland. Thus, the ‘Language Division’ sets the language standards for spelling, inflection and lexical norms of the language. Its members are representative of the research community and first language speaker in Finland and Russia, as mentioned above. Language technological descriptions and facilitation of standardization are also utilized in language education and Saami Culture Archive development at the Giellagas-Instituutti, University of Oulu, Finland (e.g. https://www.oulu.fi/giellagasinstitute/corpusproject/).

3. Skolt Sami and other Uralic languages

Skolt Sami (sms) can be set in contrast to 8 other Sami languages

(Northern Sami (sme), South Sami (sma), Pitesami (sje), Lule Sami (smj), Inari Sami (smn), Kildin Sami (sjd)) in 4 countries, Finland, Russia, Norway and Sweden. As minority languages, the Sami languages are often lumped together by less informed outsiders. Needless to say, none of the languages has majority language status that might be imposed on others at the state level, which is different from the situation of the adjacent languages Meänkieli (fit) in Sweden, and Kveen (fkv) in Norway. Here Meänkieli and Kveen are often considered to be forms of Finnish by the neighboring Finns.

The layman view of language policy affecting other Uralic languages often invokes a question of "identity consolidation" versus "divide and conquer". Both can be seen as detrimental policies. Whereas identity consolidation may be associated with mutual political entities, this same concept of entities may be the motivation for division of identity. Tundra Nenets (approx. 31,000 speakers) is spoken in a continuum spanning two Okrugs on the Arctic Ocean, while linguistic differences have not been assessed as unsurmountable, and, in fact, a mutual orthography is used.

The continuum of Komi dialects, on the other hand, has a long-standing political split between Komi-Permyak (koi) and Komi-Zyrian (kv). The rift is also apparent in the development of two separate orthographies from no later than the beginning of the 20th century. In the northern parts of the Komi Republic and adjacent areas, there is also an Iz'va (ru. Izhma) language form regarded as a dialect within the Republic yet serving as a learning medium in elementary school publications outside the Republic (cf. Rocheva, 2018).

Division of language forms into readily intelligible media means that language learning materials can be developed to address the individual users. This can be facilitated for the individuated language pairs seen in:

(1) Karelian (krl) versus Olonets-Karelian (olo);
(2) Komi-Permyak (koi) versus Komi-Zyrian (kpv) but also Komi-Izhma;
(3) Moksha (mdf) versus Erzya (myv);
(4) Eastern & Meadow Mari (mhr) versus Hill Mari (mrj), and
(5) Võro (vro) versus Seto.

It does, however, come with a cost, and this points to the logic of developing tools to utilize reusable research outcomes and results, such as is done in the Giella infrastructure. At the moment there are descriptions of various sizes for 29 Uralic languages (about a fourth of the minority languages facilitated) on the open-source Giella infrastructure. And these finite-state descriptions are made more available for research and development through open-source python libraries (Hämäläinen 2019).

4. Conclusion

The Skolt Sami language itself does not exemplify pluricentricity in its own right. The community remaining outside of Finland is welcomingly included in language development and work in the normative body. Nonetheless, Skolt Sami does provide an example of collaboration with other closely related language forms for a mutual good, i.e. the facilitation and use of multilingual and language-independent technologies. This understanding and experience is also being forwarded for other minority and not only Uralic languages, whose pluricentricity still requires evaluation.

References

Fiest, T. (2015): A Grammar of Skolt Saami. Mémoires de la Société Finno-Ougrienne, 273. Helsinki: Suomalais-Ugrilainen Seura.

Hämäläinen, Mika. (2019): UralicNLP: An NLP Library for Uralic Languages. Journal of Open Source Software, 4(37), [1345]. (https://doi.org/10.21105/joss.01345)

Hämäläinen, M. / Tarvainen, L. L. / Rueter, J. (2018): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. & Tokunaga, T. (eds.). Paris: European Language Resources Association (ELRA), p. 862-867 6 p.

Hämäläinen, Mika / Rueter, Jack (2018): Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Čibej, J., Gorjanc, V., Kosem, I. & Krek, S. (eds.). Ljubljana: Ljubljana University Press, p. 967-978 12 p.

Itkonen, T. I. (1958): Koltan- ja Kuolanlapin sanakirja – Wörterbuch des Kolta- und Kolalappischen. Lexica Societatis Fenno-Ugricae XV. Helsinki: Suomalais-Ugrilainen Seura. (https://www.sgr.fi/lexica/lexicaxv1.pdf and https://www.sgr.fi/lexica/lexicaxv2.pdf on line.)

Juutinen, Markus. (2018): *Paatsjoen Petsamon murre* [*The Paatsjoki Dialect of Petsamo*]. Oulu yliopiston Giellagas-Instituutti, Nellim. 22.9.2018. (Presentation)

Magga-Kumpulainen, Rita. (2017): Maailman ensimmäinen koltansaamelainen museo [The first Skolt Sami Museum in the World]. *Kaltio, pohjoinen* *kulttuurilehti,* Paperilehdestä 2017, №4. (accessed 24.2.2019 http://kaltio.fi/paperilehdesta/maailman-ensimmainen-kolttasaamelainen-museo/)

Miestamo, Matti. (2011): Skolt Sami: a typological profile. Journal de la Société Finno-Ougrienne 93. Helsinki: Suomalais-Ugrilainen Seura. (accessed 24.4.2019 https://www.sgr.fi/susa/93/miestamo.pdf)

Moshagen, Sjur N. / Pirinen, Tommi A. / Trosterud, Trond. 2013. Building an open-source development infrastructure for language technology projects. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16 (http://www.ep.liu.se/ecp/085/031/ecp1385031.pdf)

Moshnikoff, Jouni & Moshnikoff, Satu. (Forthcoming): Suomi-koltansaame sanakirja [Finnish-Skolt Sami Dictionary]. Giellagas Institute, Oulu & Sami Parliament of Finland, Inari.

Rocheva, Natalia. (2018): «Новый учебник Севера – Коми кыы (изьватас сёрни)» [New readers in the North – Komi language (Iz'va/Izhma speech)]. «Родные языки в условиях двуязычия» V Международная научно-практическая конференция. 25–26 октября 2018 года (г. Сыктывкар, Республика Коми). (presentation)

Rueter, Jack / Hämäläinen, Mika. (2017): Synchronized Mediawiki based analyzer dictionary development. In F. M. Tyers, M. Rießler, T. A. Pirinen & T. Trosterud (eds.), 3rd International Workshop for Computational Linguistics of Uralic Languages (IWCLUL 2017): St. Petersburg, Russia 23 – 24 January 2017., 2, Association for Computational Linguistics, Stroudsburg, pp. 1-7, International Workshop for Computational Linguistics of Uralic Languages, St. Petersburg, Russian Federation, 23/01/2017. (http://aclweb.org/anthology/W17-0601)

Sammallahti, P. (1998): The Saami Languages: An Introduction., Karasjok, Norway: Davvi Girji.

Sammallahti, Pekka & Moshnikoff, Jouni. (1991): Suomi-koltansaame sanakirja / Lää'dd-sää'm sää'nnǩe'rjj. Girjegiisá Ohcejohka.

Trosterud, Trond. (2017): *Language technology for morphologically rich langauges.* MLP 2017 : The First Workshop on Multi-Language Processing in a Globalising World. Dublin, Ireland September 4–5, 2017. (http://mlp.computing.dcu.ie/mlp2017/docs/trosterud.pdf) (presentation)

On-line databases and infrastructures

Álgu: Etymological database of the Saami languages. Online: (http://kaino.kotus.fi/algu/index.php?t=etusivu)

Giella: Giellatekno & Divvun open-source infrastructure <http://giellatekno.uit.no>

Ǩiõlljuâggtõs [Language division] of the Sami Parliament in Inari, Finland. Online through Sami Parliament in Inari, Finland (https://www.samediggi.fi/2018/03/22/nuorttsaam-kiolljuaggtos-kiorgti-jonn-askldoksannostuajas/?lang=nuo)

Max Planck Institut media archives, (http://corpus1.mpi.nl/qfs1/media-archive/dobes_data/Kola-Saami/Recordings/Language/KSDP/Skolt/-Lexicon/Vocabulary/Media/tchaeaeqcc_SKO111211_ZMN_5.wav) for example