# Samisk språkteknologi i 2021

*Sjur Nørstebø Moshagen*
*Institutt for språk og kultur*
*UiT Noregs arktiske universitet*

*Artikkelen gjev eit oversyn over samisk språkteknologi i 2021, i lag med eit stutt samandrag av historia, og går deretter inn på nokre av utfordringane i arbeidet med samisk språkteknologi og språkteknologi for minoritetar meir allment. Utfordringane blir tydeleggjorde med handfaste døme. Artikkelen blir avslutta med å peika på nokre vegar framover for å rydda dei viktigaste hindera av vegen.*

## 1. Oversyn

Arbeidet med samisk språkteknologi starta tidleg på 1990-talet, då professor Pekka Sammallahti og professor Kimmo Koskenniemi laga eit utkast til ein nordsamisk morfologisk analysator. Kring år 2000 vart arbeidet deira gjeve vidare til UiT Noregs arktiske universitet (UiT) ved professor Trond Trosterud. I etterkant av dette vart Giellatekno-gruppa[1] ved UiT oppretta for å arbeida vidare med samisk språkteknologi.

I 2004 vart Divvun-gruppa[2] starta ved det norske Sametinget, og arbeidet med å utvikla retteprogram for nord- og lulesamisk byrja for alvor i 2005, i nært samarbeid med Giellatekno-gruppa og med ein felles kodebase. Verktøya vart lanserte i 2007, og i 2008 starta arbeidet med å utvikla tilsvarande verktøy for sørsamisk. Desse vart lanserte i 2010. I 2011 vart Divvun-gruppa flytta frå det norske Sametinget til UiT og samtidig gjort permanent, i staden for å vera prosjektfinansiert. Frå då av har UiT vore senteret for forsking og utvikling av språkteknologi for samiske språk og andre urfolks- og minoritetsspråk – som til dømes færøysk og grønlandsk – med eit miljø på i snitt ti personar. Over tid har det òg vorte utvikla ein avansert og språkuavhengig infrastruktur som gjer det lett å starta arbeid med eit nytt språk og utvikla dei mest grunnleggjande verktøya ganske raskt.

1 https://giellatekno.uit.no/index.nob.html

2 https://divvun.no/

## 2. Kva finst i dag?

Over desse ikring 20 åra har det vorte utvikla ein god del verktøy. Det finst no ein infrastruktur som gjer det enkelt å definera utforminga av **tastatur**, både for datamaskiner og mobile einingar. Med denne infrastrukturen har UiT-miljøet bygt datamaskintastatur for dei fleste samiske språka, og utforminga er gjort i samarbeid med Giellagáldu[3], det allnordiske, samiske språknormeringsorganet. Det er dessutan bygt mobiltastatur for både samiske og andre språk.

**Retteprogramma** var og er ein sentral del av mykje av det arbeidet som er og framleis blir gjort ved UiT. Det starta med stavekontrollar for nord- og lulesamisk som vart bygt ut med automatisk orddeling ganske raskt, og deretter støtte for sørsamisk. Dei siste åtte åra er kjernen i retteprogramma tufta på teknologi frå Helsingfors universitet (Lindén m.fl. 2009). Verktøykassa har i seinare tid vorte vidare utvida med ein avansert grammatikkontroll basert på teknologi frå Syddansk universitet (Karlsson 1995; Bick & Didriksen 2015).

For minoritetsspråk er tilgang til gode **ordbøker** heilt sentralt, og UiT-miljøet har utvikla og tilbode elektroniske ordbøker i over ti år. Dei elektroniske ordbøkene er dei mest populære ressursane som finst i Tromsø, og har ein samla søkjefrekvens på mange tusen søk i døgnet. Det er to hovudportalar for ordbøkene, med ulik profil og ulike brukarar: NDS[4] og satni.org[5].

Språkteknologien som er utvikla ved UiT, blir òg brukt i nettbaserte **språklæringsprogram**[6] for fleire samiske og andre språk. Programma analyserer det studentane svarar, og gjev tilbakemeldingar basert på analysen.

I samarbeid med Apertium[7] er det utvikla **maskinomsetjing** for fleire par av språk som ein kan omsetja mellom (språkpar). Det einaste språkparet som er offentleg tilgjengeleg, er nordsamisk–norsk, men det blir aktivt jobba med fleire andre, mellom anna nordsamisk–lulesamisk, i samarbeid med den nordsamiske dagsavisa Ávvir[8].

Sidan 2015 har det eksistert eit **tekst-til-tale**-system for nordsamisk, utvikla i samarbeid mellom Divvun, Sametinget og Acapela[9], og sidan slutten av 2020 er det eit nytt prosjekt på gang for å gjera det same for lulesamisk. Det meste

3 https://www.giella.org

4 https://dicts.uit.no

5 http://satni.org

6 https://oahpa.no

7 https://www.apertium.org

8 https://www.avvir.no

9 https://www.acapela-group.com/

av arbeidet i det nye prosjektet blir gjort ved UiT, og all programkode skal vera open kjeldekode. Det er òg starta eksperiment med **talegjenkjenning**.

Språkteknologimiljøet ved UiT har i samarbeid med det norske Sametinget samla inn eit stort **samisk tekstkorpus**. Tekstane i korpuset er analyserte, tilgjengelege og søkbare i ein nettapp.[10] Det gjer det lett for alle å sjå til dømes korleis eit spesifikt ord blir brukt i ulike tekstar og samanhengar.

Alle verktøya som er utvikla ved Sametinget og UiT, er grammatikkbaserte og altså ikkje bygde på statistiske eller andre maskinlæringsbaserte metodar. Det er ein av hovudgrunnane til at det er mogleg å laga alle desse verktøya, sjølv for språk med få eller ingen elektroniske ressursar. Trass i at teknologi, infrastruktur og røynsle finst, er det likevel fleire nordiske minoritetsspråk og urfolksspråk som framleis har få verktøy for å kunna brukast digitalt. Det gjeld framfor alt romanispråka og dei minste samiske språka.

Tromsø-miljøet samarbeider med Uleåborg universitet, Anarâškielâ servi og Meän akateemi, i tillegg til institusjonar i Russland og Nord-Amerika, for å tilby infrastruktur og kunnskap knytt til utvikling av språkteknologi for minoritetar og urfolksspråk. Per i dag har vi grammatiske modellar for ca. 130 ulike språk og tastaturdefinisjonar for 50 språk, alle lagra i kjeldekodehandteringssystemet GitHub[11]. Nokre av modellane og tastatura er nærmast som fragment å rekna, medan andre er fullt utvikla og i dagleg bruk av tusentals språkbrukarar.

## 3. Utfordringar i 2021

Alle nordiske land har visjonar om å digitalisera samfunnet. Landa har òg språkpolitiske føringar. Den styrande tanken er at alle digitale tenester skal vera tilgjengelege for alle innbyggjarar, på innbyggjaren sitt eige språk, og minoritetsspråka skal vera samfunnsberande i sine eigne regionar, slik det er nedfelt i den nordiske språkdeklarasjonen[12]. Dette vil ikkje fungera om ikkje alle delane av den digitale infrastrukturen er lagde til rette for fleirspråklegheit og er tilpassa kvart enkelt språk.

Det er mykje som enno haltar, også for dei statsberande majoritetsspråka. Men det haltar mykje meir for minoritetsspråka, til eit punkt der det knapt går å nytta språka digitalt i visse samanhengar. I det følgjande vil eg gje nokre døme på slike bakevjer i den digitale infrastrukturen.

10 https://gtweb.uit.no/korp

11 https://github.com/giellalt

12 https://www.norden.org/no/declaration/sprakdeklarasjonen

## 3.1 Udefinerte språk

For at eit språk skal fungera i den digitale verda, må språket vera mogleg å identifisera. Det betyr at operativsystem og dataprogram veit at det eksisterer eit slikt språk, og kan slå opp kva for verktøy som finst for det spesifikke språket. For mange av språka vi arbeider med, er ikkje dette tilfellet.


*Figur 1. Slik vil ein ikkje sjå språket sitt.*

Konsekvensane av at eit språk ikkje er definert, er fleire. Brukarane vil berre sjå ein språkkode i staden for det riktige namnet på språket. Det gjer at dei ikkje kjenner att språket sitt, og verktøya blir såleis utilgjengelege. På Windows blir det uråd å nytta korrekturprogram, og det blir umogleg å nytta operativsystemtenester for språket, sjølv om ein installerer dei komponentane som trengst. I tillegg sender det eit kraftig signal til språkbrukarane om at språket deira ikkje er mykje verdt.

Den enklaste løysinga på dette er å overføra ansvaret for alle delstandardane i ISO 639 til ein felles organisasjon (dei er i dag spreidde på fleire institusjonar), til dømes Unicode[13]-organisasjonen, og deretter behandla heile ISO 639-familien på same måten som ein behandlar Unicode-standarden: Heile standarden er lagt inn i alle operativsystem, og han blir oppdatert med jamne mellomrom.

## 3.2. Tastatur

For norsk-, svensk-, dansk-, islandsk- og finskspråklege, og fleire til, er det sjølvsagt at datamaskiner, mobiltelefonar og nettbrett har tastatur for det språket brukarane nyttar. Dette er ikkje sjølvsagt for minoritetsspråka, og dei som har, har det ikkje takk vera Microsoft, Google eller Apple – med unnatak for færøysk og nordsamisk.

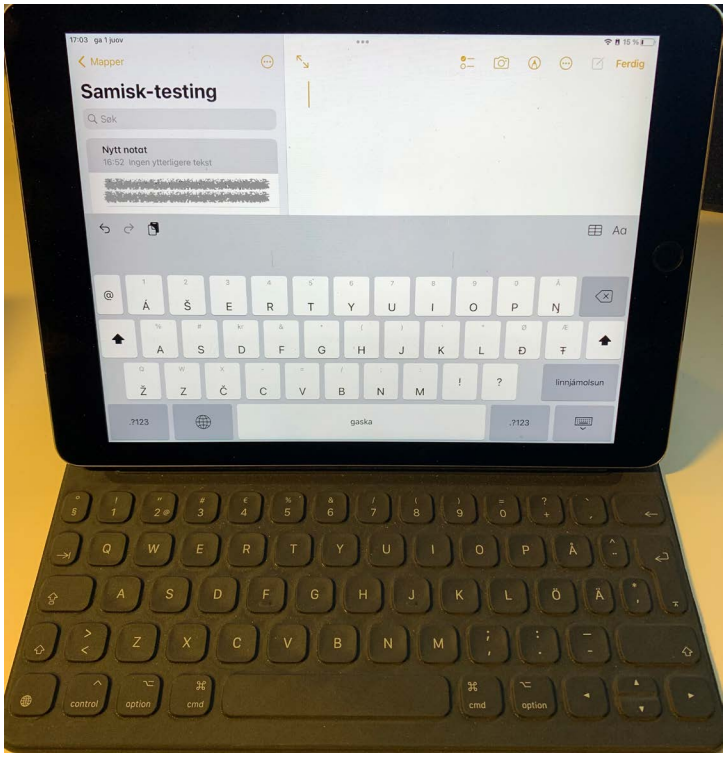
Språkteknologimiljøet ved UiT har jobba med tastaturløysingar i mange år

13 https://home.unicode.org/

og har no eit system som gjer det enkelt å laga tastatur for nye språk. Med dei verktøya som finst, har det vorte laga tastatur for dei fleste samiske språka, både mobiltastatur og datamaskintastatur. Likevel er det ein del tilfelle der ting ikkje fungerer.

Mange skular har teke i bruk nettbrett eller Chrome-bøker for elevane sine og bruker desse som hovuddatamaskin i undervisinga. Nettbretta kan sjølvsagt nytta tastatura vi har laga, men både for Android-nettbrett og for Apple sine iPad-ar er dei fysiske tastatura blokkerte for oss – vi får ikkje lov til å skriva tastaturdefinisjonar som kan bli brukte i lag med dei fysiske tastatura. Det betyr at samiske elevar heile tida må byta mellom det fysiske tastaturet (som er på norsk, svensk eller finsk) og det samiske tastaturet på skjermen. Det går – men det er ikkje slik ein vil arbeida ein heil skuledag.

Eit anna døme er Chrome-bøkene frå Google. I desse finst det ikkje innebygd støtte for eit einaste samisk språk. Det har vi gjort noko med, men tastatura er framleis variantar av majoritetsspråka. Det er Google sine rammer som gjer at vi må gå fram på denne måten for at tastatura skal fungera.




*Figur 2. Og aldri skal dei to tastatura møtast. Nordsamisk på skjermen, svensk/finsk fysisk tastatur.*

iOS-tastatura vi lagar, vil alltid vera bleike kopiar av Apple sine. Apple legg føringar og hindringar i vegen for oss som gjer at vi, sjølv med svært dyktige programmerarar, aldri kan tilby same funksjonalitet som Apple sjølv. Eit trivielt døme er emoji-tastaturet: På alle Apple sine eigne tastatur, som dei for norsk, svensk osb., ligg det ein emoji-knapp til venstre for mellomromstasten. Det gjer det lett å leggja inn ein emoji i teksten, ein funksjonalitet som så godt som alle nyttar no for tida. Men Apple hindrar oss i å utvikla denne funksjonaliteten for samiske brukarar. På grunn av restriksjonane Apple legg på oss, kan vi difor ikkje tilby ei like saumlaus og god brukaroppleving som det Apple kan, for samiske brukarar. Anten må vi klara oss utan – samiske brukarar må i staden trykkja på globusen, byta tastatur, skriva emoji og så byta tilbake – eller vi må laga vårt *eige* emoji-tastatur og halda det oppdatert heile tida. Det har vi vurdert som altfor kostbart.

Hindringane som teknologigigantane legg i vegen for oss som arbeider med språkteknologi for små språk, gjev oss ekstra kostnader og ekstra arbeid. Det blir nær umogleg å kunna tilby det folk ventar seg. Vi betalar slik ein slags «minoritetsspråksskatt» fordi dei store programvareprodusentane ikkje gjer ein skikkeleg jobb i høve til minoritetsspråka. Slik burde det ikkje vera i 2021.

### 3.3. Retteprogram

Noko av det fyrste som vart planlagt og laga då arbeidet med samisk språkteknologi byrja, var ein stavekontroll. Det er eit enkelt og etter kvart sjølvsagt verktøy for å fanga opp vanlege skrivefeil, og verktøyet er svært viktig i eit språksamfunn der mange er usikre på rettskrivinga. Dei fyrste stavekontrollane for nord- og lulesamisk kom i 2007 og har vore viktige hjelpemiddel sidan då (Antonsen & Trosterud 2020).

I mange år var den vanlegaste framgangsmåten at ein installerte eit retteprogram på si eiga datamaskin eller på ein server som IT-folka styrte med, slik at stavekontrollen fanst lokalt. Men dei siste åra har det kome ei endring, ein ny måte å laga og bruka kontorprogram på: i nettlesaren. Og plutseleg låg ikkje stavekontrollen på eiga datamaskin eller på serveren til IT-folka – han låg på serveren til Google, Microsoft eller andre IT-kjempar. Og dei vil ikkje ha verktøy og program som dei ikkje har laga sjølve. Med eitt slag var samisk språkteknologi tilbake til tida før 2007.

Dette gjeld all programvare over nettet, såkalla SaaS[14], og råkar minoritetsspråk mykje hardare enn majoritetsspråk. Med SaaS er det ikkje lenger mogleg

14 SaaS = Software as a Service

å installera eigne tillegg, slik som ein stavekontroll, og om det er, er både funksjonaliteten og brukargrensesnittet kraftig redusert.

Det er ikkje umogleg å få til ei fungerande løysing som tek vare på personvern og datatryggleik. Til dømes kunne dei store IT-selskapa ha eigne, sertifiserte koplingspunkt som andre skulle kunna nytta etter avtale, samtidig som brukarane skulle bli informerte om kva organisasjon retteprogramma kjem frå. Men ingen av dei store selskapa har så langt gjort nokre forsøk i ei slik retning.

## 4. Utfordringar framover

UiT-miljøet ved Divvun- og Giellatekno-gruppene har vist at det er mogleg å byggja grammatikkbasert språkteknologi for minoritetsspråk og for språk med små eller ingen elektroniske ressursar. Det går an å laga funksjonelle og nyttige verktøy for alle språksamfunn i ein elektronisk kvardag. Dette er dessutan heilt naudsynt for at språka skal halda fram med å vera bruksspråk i sine eigne samfunn. Det kjem ikkje av seg sjølv, og arbeidet som har vorte gjort, er eit resultat av ei målretta satsing av det norske Sametinget og det norske Kommunal- og distriktsdepartementet (KDD) over mange år. Utan innsatsen deira ville vi ikkje hatt dei verktøya vi trass alt har i dag.

Samtidig ser vi at digitaliseringa berre går raskare og raskare, og frå andre urfolkssamfunn og minoritetar rundt om i verda veit vi at trykket mot dei små språka berre aukar. Mange språk kjem til å gå or bruk og døy ut dette hundreåret.

Det neste store steget innanfor språkteknologi ser vi allereie: digitale assistentar som ein kan prata med og få munnlege svar tilbake frå. Dei er enno ganske enkle, men fungerer likevel godt for dei domena dei kan handtera. Hovudproblemet er at dei berre finst for dei største og økonomisk mest interessante språka.

Og slik kjem det i all hovudsak til å bli verande. Ein kan ikkje venta at Google, Microsoft, Apple eller andre tek på seg å laga språkteknologi for 7000 språk, heller ikkje for dei omkring 3500 språka som framleis vil vera i dagleg bruk om hundre år (Crystal 2000). Derimot veit vi at grupper som språkteknologimiljøet ved UiT vil arbeida for å gje verktøy til utsette språksamfunn.

Utfordringane knytte til slikt arbeid er både teknologiske og språklege.

### 4.1. Teknologiske utfordringar

Språkteknologi, særleg avansert språkteknologi som maskinomsetjing og digitale assistentar, er viktige drivkrefter i den digitale økonomien og er ikkje noko som dei store selskapa lett opnar opp. Like fullt er det det dei må. Dei må gjera det mogleg for andre partar enn dei sjølve å utvikla og levera teknologiske

løysingar på dei språka dei sjølve ikkje tek ansvar for. Det er mest synleg med taleteknologi – skal ein kunna prata samisk med bilen sin i framtida, må nokon laga eit samisk system for talegjenkjenning og talesyntese – men gjeld heile den språklege infrastrukturen, frå menytekstar til Siri. Det må vera mogleg for kvart språksamfunn å sjølve ta styringa over teknologien som gjeld deira eige språk. Om ikkje Google, Apple, Microsoft eller dei andre store vil ta ansvar for eit språk – og det er mange språk dei ikkje kan ta ansvar for – så skal dei ikkje samtidig hindra andre i å gjera det.

Det betyr i praksis at alle delar av operativsystema og tenestene dei leverer som gjeld naturlege språk, må opnast opp. Dei treng ikkje opna opp den underliggjande teknologien, men dei må gjera det mogleg for andre å plugga inn sin eigen teknologi for språk som dei sjølve ikkje har teknologi for. Alle operativsystem skil i dag mellom språkspesifikke delar og språkuavhengige komponentar. Detaljane varierer sjølvsagt, men skiljet er der. Det betyr at grunnlaget for å gjera dei språkspesifikke delane tilgjengelege eller opne finst. Det vil krevja ein del arbeid, men det er eit naudsynt arbeid om alle språksamfunn skal kunna bli ein del av det digitale samfunnet på like vilkår. Når eit slikt skilje er gjennomført, dei språkspesifikke delane er opne og andre har byrja å laga slike komponentar for nye språk, kan alle dei store systemprodusentane distribuera språkpakker frå tredjepartar gjennom sine eigne programvarebutikkar.

### 4.2. Språklege utfordringar

Dei språklege utfordringane handlar om ressursar og språkutvikling. Teknologien vi nyttar ved UiT, er grammatikkbasert, og det trengst folk som kan grammatikken til språka og kan skriva han om til noko som datamaskinene forstår. Det trengst òg arbeid med terminologi og leksikografi, og det trengst arbeid med skriftspråksnormering. Dette arbeidet er ein viktig del av det å ta tilbake språka og gjera dei samfunnsberande, og arbeidet er heilt naudsynt om ein skal oppfylla intensjonen bak den nordiske språkdeklarasjonen.

## 5. Oppsummering

Eg har i denne artikkelen gjeve eit oversyn over det arbeidet som har vorte gjort for samisk og andre minoritets- og urfolksspråk ved UiT. Gjennom arbeidet vårt har vi vist at det er mogleg å laga gode verktøy for språk med få ressursar og kompleks grammatikk. Samtidig har eg peikt på utfordringar knytte til det å få tilgang til verktøya i ein del samanhengar og ulike hindringar som dei store teknologiselskapa legg i vegen for dei små språksamfunna. Eg har samtidig prøvd å peika på moglege vegar framover. Målet må vera at kvart samfunn sjølv skal kunna ta ansvar for sin eigen språkteknologi, i samarbeid

med partar som dei sjølve kan velja. Berre slik kan vi byggja ei digital framtid som inkluderer alle.

## Summary

In this article I give an overview over what has been achieved by the Divvun and Giellatekno groups at *UiT The Arctic University of Norway* over the 20 years since the work started, both in terms of concrete products and tools, but also in infrastructure development enabling new language communities to enter the digital train. I then discuss some concrete issues we have met concerning the IT giants' lack of attention regarding the consequences of their actions towards minority and indigenous language communities. I end the article by pointing out how these issues could be solved in general, by opening up the field of language technology components to third parties and let them offer language packs for languages not supported by the tech giants, through the app stores of the platforms. These steps are necessary to minimise language death as much as possible.

## Forfattar

**Sjur Nørstebø Moshagen** har arbeidd med språkteknologi i både industri og akademia i ikring 30 år og har vore leiar for Divvun-gruppa ved UiT sidan starten i 2004.

## Litteratur

Antonsen, Lene & Trond Trosterud, 2020: Med et tastetrykk. Bruk av digitale ressurser for samiske språk. / Boallobeavddi bokte. Sámegielaid digitála resurssaid geavaheapmi. I: *Sámi logut muitalit 13. Čielggaduvvon sámi statistihkka 2019 / Samiske tall forteller 13. Kommentert samisk statistikk 2020*. Sámi allaskuvla.

Bick, Eckhard & Tino Didriksen, 2015: CG-3 - Beyond Classical Constraint Grammar. I: Beáta Megyesi (red.): *Proceedings of NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. s. 31–39. Linköping: LiU Electronic Press.

Crystal, David, 2000: *Language Death*. Cambridge University Press, Cambridge.

Karlsson, F., A. Voutilainen, J. Heikkila, & A. Anttila, 1995: *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin.

Lindén, Krister Miikka Silfverberg & Tommi Pirinen, 2009: HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. I: Mahlow, Cerstin & Michael Piotrowski: *State of the Art in Computational Morphology*, s 28–47. Springer Berlin Heidelberg.

## Nøkkelord

språkteknologi, minoritetsspråk, språkpolitikk