Med et tastetrykk

Bruk av digitale ressurser for samiske språk

Lene Antonsen og Trond Trosterud, UiT Norges arktiske universitet

Sammendrag

Digital kommunikasjonen mellom myndigheter og brukere blir stadig mer utbredt, og mange offentlige etater forventer at brukerne skal ta kontakt via digitale kanaler. Det arbeides nå for at også samiske brukere skal kunne bruke sitt eget språk i slik kommunikasjon. I denne artikkelen ser vi på hva som finnes av skriveverktøy og annen programvare til hjelp i skrive- og leseprosessen for samiske språk. I dag står de samiske språka relativt sterkt i digital sammenheng, men det var langt fra opplagt at det skulle gå slik. For nordsamisk skjedde innføringa av ei ny felles rettskriving, innføring av samisk skriftspråk i offentlig sammenheng og begynnelsen på dataalderen omtrent samtidig. Denne artikkelen viser hvorfor de samiske språka trass i et vanskelig utgangspunkt står såpass sterkt i digital sammenheng som de gjør, og hvilke konsekvenser dette har fått for den samiske språksituasjonen. Vi har også henta inn statistikk over bruk av en del digitale samiske språkverktøy.

Samisk språkteknologi endrer folks atferd. Internasjonal standardisering gjorde det mulig å få samisk på internett på 1990-tallet. Samiske retteprogram har blitt en del av hverdagen for de aller fleste som skriver samisk. Innføringa av samiske tastatur på mobiltelefonen økte kvaliteten på samisk tekst på Facebook drastisk. Da maskinoversetting ble tilgjengelig, forsvant store deler av diskusjonen om hvorvidt det skulle være lov til å skrive på samisk i nettdiskusjoner. Nettordbøker med grammatikk og lenke til autentisk samisk tekst går et langt steg i å kompensere for den svake posisjonen samiske språk har i samfunnet som helhet.

Oversetting av tekst til samisk utgjør en flaskehals i svært mange sammenhenger, særlig når det gjelder skolebøker og sakprosa, så mer effektive rutiner for oversetting vil ha stor innflytelse på tilgang til samisk tekst. Bruk av oversettingsminne og automatisk oppslag av fagtermer vil også gjøre det lettere for oversetterne å få en mer konsistent terminologi. Til slutt drøfter vi utviklinga av språkteknologiske løsninger framover, og hva det innebærer for arbeidet med å styrke bruk av samiske språk på ulike samfunnsområder.

1 Innledning

For majoritetsspråkbrukere i Norden har det alltid vært en selvfølge å kommunisere skriftlig med offentlige myndigheter. De siste tiåra har kommunikasjonen gått mer og mer over til digital form, og mange offentlige etater forventer at brukerne skal ta kontakt via digitale kanaler. Det arbeides nå for at også samiske brukere skal kunne bruke sitt eget språk i slik kommunikasjon. Siden samisk ikke har vært brukt i kommunikasjon med det offentlige tidligere, trenger man ord og terminologi for denne kontakten. Dessuten har mange samisktalende hatt liten eller i verste fall ingen skriveopplæring på skolen, og for alle samisktalende gjelder det at de i liten grad blir eksponert for det skrevne ord på samisk. Satsninga på digitalisering også for samiske språk kan på lengre sikt bøte på dette, men for å komme så langt må en del grunnsteiner være lagt. Samiske tastatur, retteprogram og andre hjelpemidler må være tilgjengelig, og det samme må kunnskapen om og bruken av dem være. Når det gjelder å skrive på samisk på internettsider og i sosiale medier, er det et viktig hinder at ikke alle lesere behersker språket, og da er maskinoversetting fra samisk til majoritetsspråket et relevant verktøy.

Digitaliseringsprosessen innebærer en kontinuerlig utvikling av nye verktøy, som intelligente søkemotorer på internett, oversettingsprogram mellom flere språk, maskinell opplesing av tekst, program som kan ta imot taleinstrukser, og så videre. For at minoritetsspråk skal kunne overleve i en hverdag hvor vi bruker mer og mer av slike tjenester, er det nødvendig at slike program også er utvikla for disse minoritetsspråka.

I denne artikkelen vil vi se på hva som finnes av skriveverktøy og annen programvare til hjelp i skrive- og leseprosessen for samiske språk. I den grad vi har fått tilgang til logger eller statistiske opplysninger, vil vi også se på bruken av programma.

2 Bakgrunn

2.1 Digitalisering og språk

Digitalisering er en prosess som *konverterer informasjon til et digitalt format*, dvs. maskinlesbar informasjon i et format bestående av tall (null eller en). I denne artikkelen vil vi begrense oss til digitalisering av samiskspråklige data, og dataprogramma som blir brukt for å prosessere slike data, og som er åpent tilgjengelige for samisktalende, med vekt på situasjonen i Norge.

I dag står de samiske språka relativt sterkt i digital sammenheng, men det var langt fra opplagt at det skulle bli slik. For nordsamisk skjedde innføringa av ei ny felles rettskriving, innføring av samisk skriftspråk i offentlig sammenheng, og begynnelsen på dataalderen omtrent samtidig (ny rettskriving i 1979, og de andre prosessene fra og med 1980-talet). De samiske språka gikk inn i dataalderen med bokstaver som ikke var brukt i alfabeta til flertallsspråka: sørsamisk *ï*, lulesamisk *á, ŋ*, nordsamisk *á, č, đ, ŋ, š, ŧ, ž*, enaresamisk *â, á, č, đ, ŋ, š, ž*, skoltesamisk *â, å, č, ʒ, ǯ, ŋ, đ, ǧ, g, ǩ, ŋ, õ, š, ž, ´* og kildinsamisk *ā, ä, ē, ū, ӆ, ӎ, ӈ, ӌ, ō, p, ӯ, ө, э, ӭ, ӧ, ӓ, ҧ*. Mengden av ekstrabokstaver øker jo lenger nordøstover vi kommer, det samme gjør behovet for tilpassede tastatur og tegnsett. Også for språkteknologiske verktøy var utgangspunktet vanskelig. Måten å lage ordretteprogram var utvikla for engelsk og andre vesteuropeiske språk, og de samiske språka har en grammatisk struktur som gjør at den same måten å lage program på passer dårlig for samisk. Mye tyda derfor på at de samiske språka skulle få store problemer med overgangen til digital språkbehandling.

Med denne artikkelen vil vi vise hvorfor de samiske språka trass i et vanskelig utgangspunkt står såpass sterkt i digital sammenheng som de gjør, og hvilke konsekvenser dette har fått for den samiske språksituasjonen. Vi har også henta inn statistikk over bruk av en del digitale språkverktøy. Til slutt vil vi drøfte utviklinga av språkteknologiske løsninger framover, belyse noen mangler og se på hva det innebærer for arbeidet med samisk språk.

2.2 Språkteknologi

I utgangspunktet er språkteknologi all teknologi som kan brukes for å behandle språk. I dag blir termen brukt om *dataprogrammas evne til å analysere, produsere, endre eller respondere på menneskelig tekst og tale*. Språkteknologi er en viktig del av teknologifeltet. Vi mennesker bruker naturlig språk når vi interagerer både med hverandre og med datamaskinene, og det å gjøre maskinene i stand til å behandle naturlig språk griper dermed inn i store deler av programvareutviklinga. Nettopp derfor er også språkteknologien djupt integrert i de ulike dataløsningene, og de ligger også an til å få ei viktig rolle i det nye *Tingenes internett*, der vi kan kommunisere med høyttalerne, kjøleskapene, ytterdørene og bilene våre. Spørsmålet er om vi da kan bruke samisk?

2.3 Åpen kildekode

Programvareutvikling foregår langs to ulike framgangsmåter: *lukka* og *åpen* kildekode. Åpen kildekode skiller seg fra lukka ved ikke bare å gi tilgang til den ferdige programfila, men også til filene som blei brukt for å lage den. Med en slik tilgang kan man lage sin egen versjon av programmet og for eksempel tilpasse det til egne formål. Åpen kildekode er viktig av to grunner:

Hvis språkmodellkoden er kontrollert av programvarehuset som lager brukerprogramma den skal integreres i, risikerer språksamfunnet (dvs. de som bruker språket) at de lingvistiske ressursene språkbrukerne har vært med på å lage, bare dukker opp i programma laget av dette programvarehuset, og ikke i andre. I verste fall går ressursene tapt hvis programma som de er integrert i, går ut av bruk. Lukka kildekode vil også stenge språkmodellene ute fra programvare som har som krav at alle bestanddelene skal være åpen kildekode, for eksempel i akademiske miljøer eller program laga på dugnad. For store språksamfunn blir dette løst ved at det lages flere konkurrerende språkmodeller, noen åpne og andre lukka. Små språksamfunn har ikke ressurser til dette, og derfor er åpen kildekode viktig for samiske språk. I kapittel 4.3 er det eksempler på at åpen kildekode gir flere aktører muligheter til å lage ordbøker og ordboksapper.

3 Samisk språkteknologi

Her presenterer vi eksisterende infrastruktur og programvare for behandling av samiske språk. Det største miljøet som arbeider med dette, er ved UiT Norges arktiske universitet. Vi er selv en del av dette miljøet, og dette vil også bli gjenspeilet i presentasjonen vår. Vi vil likevel også legge vekt på å presentere andre miljø og initiativ.

3.1 Oppstarten for samisk språkteknologi

Arbeidet med å bygge opp språkteknologi for samiske språk startet i 2001 ved UiT, de første åra bare med éi stilling, men fra 2008 var det tre faste stillinger, som utgjorde forskningsgruppa *Giellatekno*. Det norske sametinget etablerte i 2004 *Divvun*-gruppa, som hadde som formål å bygge ordretteprogram for de tre offisielle samiske språka i Norge. Divvun-gruppa samarbeida fra begynnelsen av tett med UiT og blei i 2011 også flytta formelt fra Sametinget til UiT, med finansiering fra Kommunal- og moderniseringsdepartementet. Gruppene har sammen bygd språkteknologiske ressurser for flere samiske språk.

3.2 Samiske kodetabeller

En *kodetabell* er i denne sammenhengen ei liste av bokstaver og andre tegn der hvert tegn har en unik tallverdi, slik at datamaskinen skal være i stand til å prosessere det. I løpet av 80-tallet vokste det fram ulike ad hoc-kodetabeller for nordsamisk, der bokstaver brukt i andre europeiske språk blei erstattet med samiske bokstaver. Dataløsninga som gikk under navnet *Winsam*, var for eksempel avhengig av egne nordsamiske *typesnitt* (også kalt fonter), der bokstavene i den vesteuropeiske kodetabellen blei erstattet slik: *ç → č, ó → š, ð → đ, þ → t og ñ → ŋ*. Kodetabellene som dermed

oppsto, hadde ingen offisiell status, men var avhengige av at brukerne installerte typesnitt der bokstaven *Ç* blei gjengitt som *Č*, osv. Så lenge datamaskina blei brukt som ei skrivemaskin med printer i samme rom, gikk dette bra, men bare det å sende et manuskript i diskettform f.eks. til en forlegger som ikke hadde de rette samiske typesnitta, førte til at den samiske teksten ikke kunne trykkes. Den nordsamiske ortografien fra 1979 hadde endelig fjerna grensene som delte opp nordsamisk, men nå var det samiske språksamfunnet på nytt i en situasjon der riksgrensa blei et hinder: På den andre sida av grensa hadde de andre kodetabeller. Å publisere på internett (da det kom) var heller ikke mulig.

Det var ikke bare samisk som hadde vanskeligheter med å representere bokstavene sine på 1980-tallet. De store dataprodusentene og det internasjonale standardiseringsorganet ISO slo i 1993 sammen de respektive tegnsettstandardene sine til en felles standard for alle bokstaver for alle levende (og etter hvert også alle utdødde) språk, *Unicode*. Fra og med 1993 var det klart at alle problemer med bokstaver i prinsippet var løst. Det skulle likevel gå godt over ti år før Unicode var implementert i nye datamaskiner, og fremdeles (i 2020) kan samiske tegn dukke opp som spørsmålstegn, f.eks. i NRK Sápmi sin app for smarttelefoner.

3.3 Regelbasert språkteknologi

Det finnes ulike språkteknologiske metoder, og ressursene som finnes for språket det gjelder, er avgjørende for valg av teknologi. Det finnes ikke store nok tekstsamlinger til at de inneholder alle samiske ord, heller ikke for nordsamisk, selv om tilgjengelig tekst for nordsamisk er 30 ganger større enn for andre samiske språk, se kap. 4.4. Løsningen er derfor å generere alle orda.

Det er ikke nok å lage en liste over alle grunnformene og automatisk legge til forskjellige bøyningsendinger til ordstammene, fordi bøyning av samiske ord ofte medfører endring også av bokstaver inne i stammen av ordet, som i orda *g**oa**hti* - *go**đ**iide* og *g**åe**tie* - *g**öö**tide* (gamme/hus i grunnform og illativ flertall, på henholdsvis nordsamisk og sørsamisk). Derfor har man heller valgt å lage en datalingvistisk modell for hvert samisk språk. Modellen inneholder lister av ord på språket, og den genererer alle bøyningsformer og ordavledninger for disse orda, i tillegg til å lage alle mulige sammensetninger med orda. I tillegg til selve ordet gir modellen også grammatisk informasjon om hvert ord. Slik kan man for eksempel produsere orda som skal være med i et ordretteprogram. Ord som ikke kan genereres for det språket modellen er laga for, blir merka med en rød strek. Deretter genererer modellen fem forslag til riktige skrevne ord, som likner på ordet som ikke blei godkjent.

I og med at det kan generere orda, kan programmet også analysere dem, dvs. fortelle hva som er grunnformen, og hvilken bøyningsform (ev. ordavledning eller sammensetning) ordet har.

Dessuten har man et program som kan analysere hele setninga og fortelle hvilken funksjon hvert ord har i forhold til de andre orda i setninga, f.eks. om ordet *beatnaga* (= hund) står som eier av noe eller som objekt i setninga. Disse to programma sammen er grunnlaget for oversettingssystemet som beskrives i kapittel 4.5.1. Begge programma består av regler formulert av lingvister. Derfor kalles dette *regelbasert språkteknologi*, i motsetning til *statistisk basert språkteknologi*. (Les mer om regelbasert språkteknologi for samiske språk i Antonsen & Trosterud 2010 og Antonsen 2018.)

4 Språkteknologiske verktøy og bruken av dem

Det aller meste av språkteknologiske verktøy for samiske språk er laga ved UiT, i samarbeid med brukermiljø og andre aktører. I dette kapitlet gir vi statistikk både for verktøy som kan lastes ned, og for verktøy som brukes online. For den første gruppa har vi bare loggdata for hvor mange ganger programma er lasta ned, mens vi for den andre gruppa til en viss grad også har loggdata for bruken. Vi ser på disse loggene og drøfter implikasjonene av det de viser.

4.1 Tastatur

For å skrive på datamaskin er det ikke nok å ha bokstaver, det må også finnes et tastaturoppsett for å skrive dem inn. Å lage samiske tastatur har vært en viktig del av det samiske språkteknologiske arbeidet fra starten av.

Da personlige datamaskiner kom i bruk på 1980-tallet, kom de med tastatur som kunne produsere tekst på engelsk, norsk, svensk, finsk, tysk og fransk. Alle samiske skriftspråk fra og med nordsamisk og østover hadde en ortografisk standard med bokstaver som ikke kunne bli produsert med disse tastaturene, og utviklinga tok dermed et steg bakover fra de manuelle skrivemaskinene. Selv om det hadde gått sakte å skrive f.eks. nordsamisk med manuelle skrivemaskiner, hadde det likevel vært mulig å få det til. Bokstaven *Š* kunne f.eks. bli skrevet som en *S* med både akutt og grav aksent.

Med de nye datamaskinene var dette ikke lenger mulig, og det måtte skapes nye løsninger for de samiske bokstavene. Selv om det i og for seg kunne ha oppstått en standardisert samisk tastaturlayout, laga ulike aktører hver sin standard. På midten av 1990-tallet hadde to norske og en finsk aktør laga samiske tastaturer, alle med de samiske bokstavene på ulike posisjoner.
I 1997 foreslo det fellesnordiske *Samisk datautvalg* en ny tastaturlayout for nord-, enare- og skolte-samisk og distribuerte den sammen med de nye samiske kodetabellene. Denne tastaturlayouten blei marginalt endret av Samisk parlamentarisk råd (bokstaven *Ŧ* fikk posisjonen til *Y* heller enn *AltGr*

T, selv om *Y* var vanligere enn *Ŧ* i samisk tekst), og deretter vedtatt som standard av de nordiske standardiseringsorgana. Den har siden vært en fast del av operativsystema for Macintosh og Windows. Også Linux-distribusjonene kommer med samiske tastaturer ferdig installert.

Det har oppstått et nytt problem når det gjelder de fysiske tastaturene. Verdens nest største datamaskinprodusent, HP, produserer fra og med 2017 store deler av modellene sine med det såkalte ANSI-tastaturet, der datamaskiner i Europa ellers blir solgt med ISO-tastatur. Forskjellen dem i mellom er at på ANSI-tastaturet er tasten for nordsamisk *Ž* flyttet fra venstre til høyre side, og tasten for nordsamisk *Đ* er plassert en rad opp og en posisjon til høyre. For alle som skriver samisk er dette ei stor ulempe, og ei påminning om at framskritt for minoritetsspråk trenger kontinuerlig oppfølging.

Med innføringa av smarttelefoner fra og med 2007 starta tastaturproblemet på nytt. Telefonene var i utgangspunktet lukka system, bare produsentene kunne legge til støtte for ulike språk, og de produserte i utgangspunktet bare tastaturer for en handfull språk. Et godt eksempel er ukrainsk, som med 60 millioner talere ikke fikk tastaturlayout for iPhone, slik at det vokste opp en illegal industri for å endre telefonene, med tap av garantirett som resultat. I 2014 åpna Apple for at eksterne programvarehus kunne legge til tastatur for iPhone. Det første nordsamiske tastaturet kom samme år, og fra og med 2016 har Divvun og Giellatekno tilbudt tastatur og stavekontroll for både Apple- og Android-mobiltelefoner. I perioden desember 2014–15.5.2020 har det samiske tastaturet for iOS (iPhone og iPad) blitt lasta ned via AppStore 49 200 ganger, og det tilsvarende Android-tastaturet via Google Play 6630 ganger. 44 600 av nedlastingene blei utført i Norge, 2120 i Finland, 1060 i Sverige, og resten primært fra Europa. Fra og med 2014 finnes det også et iPhone-tastatur for nordsamisk laga av Tim Valio[1].

Gitt at de yngste og aller eldste aldersgruppene ikke bruker mobiltelefon, er det samlede nedlastingstallet for mobiltastaturer (56 000) mer enn tre ganger så stort som tallet på samiskspråklige mobiltelefonbrukere[2]. Selv om mobilbrukerne kan ha skifta mobil eller oppdatert tastatur flere ganger i løpet av perioden, viser nedlastingstalla at den alt dominerende delen av det samiske språksamfunnet har tilgang til samiske bokstaver på mobiltelefonen sin. Tilgangen til samiske tastaturer for mobiltelefoner har hatt innflytelse på bruken av samisk i sosiale medier. Samiske språk har helt fra starten av stått sterkt på Facebook. En av grunnene kan være at det fra 1990-tallet fram til Facebook blei vanlig, eksisterte et tilsvarende nettsamfunn for samer, nemlig *Saminet*. Facebook fikk også relativt tidlig et brukergrensesnitt på samisk, dvs. at brukerne kan

[1] https://www.nrk.no/sapmi/endelig-mulig-a-skrive-pa-samisk-pa-iphone-1.12041195

[2] Anslått til 20 000 av omtrent 29 000 samisktalende i alt (jf. tabell 1 for oversikt og referanser).

velge å ha menyer og overskrifter på samisk. I 2013 var 35 % av tekst på nordsamisk på Facebooks diskusjonssider skrevet uten samiske bokstaver. I 2019 var denne prosenten nede i 5 %. Det som hadde skjedd i mellomtida, var at smarttelefonene fra og med 2014 hadde fått tilgang til samiske tastatur, og at bortimot hele språksamfunnet hadde lasta dem ned.

4.2 Ordretteprogram

De første versjonene av ordretteprogrammet Divvun blei lansert seinhøstes 2007 for nordsamisk og lulesamisk. For sørsamisk blei den første versjonen ferdig i 2010. Programma har fulgt de samiske normeringsrådenes[3] vedtak, i tillegg til ordbøker og grammatikkbøker og generalisering av skriftspråkets prinsipper (jf. Antonsen 2013). Etter at Sámi Giellagáldu[4] blei opprettet i 2015, er det dette organet som normerer nord-, lule-, sør-, enare- og skoltesamisk. Ordretteprogramma kommer stadig ut i nye versjoner som er tilpasset nye versjoner av operativsystema for PC og Mac, mens det lingvistiske innholdet ikke oppdateres like ofte. Det er også laga et ordretteprogram for enaresamisk[5] (Morottaja mfl. 2018), og beta-versjoner[6] for skoltesamisk[7] (jf. Rueter & Hämäläinen 2020) og pitesamisk[8].

Ordretteprogrammet ser bare enkeltord og ikke sammenhengen de står i, og dermed kan ikke programmet vurdere om valget av ord er riktig, eller om det er brukt riktig bøyningsform av ordet i den aktuelle sammenhengen. En evaluering av den nordsamiske versjonen av programmet viser at det likevel finner omtrent 80 % av skrivefeilene i tekster skrevet av morsmålstalere, og at det gir riktig forslag til retting i 82 % av tilfella (Antonsen 2013).

Divvun-programma kan lastes ned fra internett, og installeringspakkene kan også lastes ned og distribueres til flere datamaskiner innafor samme institusjon. Siden 2016 har det dessuten vært mulig å bruke programmet online, noe som har gjort det mulig å bruke programmet når det av forskjellige grunner ikke er mulig å installere programmet på egen datamaskin. Fra og med 2019 har Divvun lansert *Divvun Installer*, der brukeren laster ned installasjonsprogrammet, som deretter sørger for at stavekontrollen til enhver tid er oppdatert.

---

[3] Sámi Giellaráđđi og Sámi giellalávdegoddi

[4] https://www.giella.org

[5] Ordretteprogrammet for enaresamisk var et samarbeidsprosjekt mellom Giellatekno ved UiT og Anarâškielâ servi.
[6] Beta-versjon betyr at programmet enda ikke inneholder alle ord og bøyninger.
[7] Ordretteprogrammet for skoltesamisk var et samarbeidsprosjekt mellom universitetet i Oulu, Giellatekno ved UiT og Jack Rueter ved Helsingfors universitet.
[8] Arbeidet blei starta av Ann-Charlotte Sjaggo og Trond Trosterud (Sjaggo og Trosterud 2015) og deretter fortsatt av Joshua Wilbur (universitetet i Freiburg, senere ved universitetet i Tartu).

Sett under ett blei alle Divvun-programma for de samiske språka i Norge lasta ned 20 555 ganger i perioden 12.4.2010–28.3.2017[9]. De fordelte seg på de ulike språka på følgende måte: retteprogram for nordsamisk 2983 nedlastinger, for lulesamisk 576, for sørsamisk 493 og felles retteprogram for alle tre språk 16 503. Hvis vi i stedet ser på fordelingene landa imellom, ser vi at retteprogramma har blitt lasta ned 10 819, 2577 og 1541 ganger fra henholdsvis Norge, Sverige og Finland, og 5618 ganger fra andre land[10]. I prosent tilsvarer dette at 52 %, 44 % og 64 % av de samisktalende i hvert av de tre landa har lasta ned programmet. På den ene sida har nok mange språkbrukere lasta ned programmet flere ganger, men på den andre sida er det en relativt stor del av de samisktalende som ikke skriver samisk på datamaskin i det hele tatt[11], og vi konkluderer med at de aller fleste som skriver på samisk, har tilgang til samiske retteprogram. Siden 2016 har det også vært mulig å bruke Divvun-programmet online[12] og som retteprogram på systemnivå[13].

I februar 2020 lanserte Divvun en tidlig versjon av en grammatikkontroll for nordsamisk, dvs. et program som ikke bare retter skrivefeil, men også feil bruk av ord som er skrevet riktig, og feil samskriving eller særskriving av ord (Wiechetek mfl. 2019).

4.3 Ordbøker

Ordbøker er spesielt viktige for minoritetsspråkstalere, som ikke på langt nær blir eksponert for språket sitt på alle samfunnsområder slik som majoritetsspråkets talere.

4.3.1 Ordbøker med bøyningsformer

Giellatekno sine NDS-ordbøker (NDS = *Neahttadigisánit*, «Nettordbok») består av søkeord i grunnform og en eller flere oversettelser. I norsk–nordsamisk ordbok er det nesten 3000 setnings-eksempler, for andre språkpar er det færre. NDS-ordbøkene kan grammatikk, dvs. at når brukeren skriver inn et bøyd ord, gir ordboka grammatisk informasjon om søkeordet, i tillegg til grunnform og oversettelse. Ved å klikke på det samiske ordet får man en tabell med bøyningsformer av ordet,

[9] Statistikken bak dette avsnittet er henta fra http://divvun.no/Download_log.html.
[10] Landet utafor det samiske området med flest nedlastinger var USA med 2267, deretter kom Kina med 543 og Tyskland med 331.
[11] Vi anslår at godt over en tredel av alle samisktalende aldri skriver samisk på datamaskinen, fordi de er for unge eller for gamle eller skriver bare på majoritetsspråket. Dette stemmer godt med resultata i Melhus og Broderstad 2020:26ff, som viser at 37,7 % og 31,3 % (dvs. svaralternativa *svært bra, nokså bra, med anstrengelse*) av de samiske respondentene kan henholdsvis snakke (fig. 9) og skrive (fig. 11) samisk, dvs. at 83 % av de samiskspråklige respondentene også kan skrive samisk. Ser vi på aldersgruppene som i større grad skriver på datamaskin (77 % av befolkninga i Finnmark er 10–70 år, jf. https://www.ssb.no/statbank/table/07459/), tyder resultata deres på at 64 % av den samiskspråklige befolkninga kan skrive samisk på datamaskin.
[12] http://divvun.no/korrektur/speller-demo.html
[13] Stavekontrollene blir laget med en åpen kompilator (HFST). Det har gjort det mulig å integrere dem i LibreOffice, og å utvide funksjonaliteten til programmene.

som blir generert der og da ved hjelp av språkmodellen som er beskrevet i kapittel 3.3. Det er også mulig å velge å søke etter ordet i tekster via grensesnittet *Korp*, se kapittel 4.4, og dermed få tilgang til autentiske språkeksempel.

Ordpara for ordboka nordsamisk–norsk bygger opprinnelig på Nils Jernslettens *Álgosátnegirji*, som har 4000 ordpar, men seinere har det blitt lagt til ordpar fra flere kilder, bl.a. blei det lagt til 15 000 ordpar fra ei tekstsamling bestående av administrative tekster oversatt fra norsk til nordsamisk[14]. Ordbøkene for nordsamisk–finsk er basert på leksikalsk materiale fra *Institutionen för de inhemska språken* i Finland, som igjen bygger på Konrad Nielsens ordbøker (1932–1962). I tillegg er det til begge ordbøkene lagt til noen tusen av de vanligste orda på majoritetsspråka, sammen med oversettelse.

Ordbøkene for sørsamisk og norsk bygger på materiale utarbeida av Albert Jåma og Tove Brustad i Hemnes sameforening[15]. Alle disse ordbøkene har fått mange ordpar fra arbeidet med språklæringsprogramma *Oahpa*. For sørsamisk var språksenteret *Aajege* i Røros med i arbeidet.

Den enaresamiske ordboka inneholder først og fremst Valtonen & Olthuis' ordbok (2016). Ordpara i den skoltesamiske ordboka kommer fra Sammallahti & Mosnikoffs ordbok (1991). Den pitesamiske ordboka bygger på *Pitesamisk ordbok* (Wilbur 2016). Filene med ordpar til de nevnte ordbøkene er nedlastbare på internett under lisensen CC-BY. Det betyr at også andre aktører kan bruke ordboksmaterialet, noe vi kommer tilbake til i kapittel 4.3.2.

[14] Arbeidet med dette ble finansiert av det daværende Fornyings-, administrasjons- og kirkedepartementet.
[15] http://www.ruovatsijte.no/gaerjiste-vaalteme_2001.pdf

Tabell 1. Bruk av NDS-ordbøkene i perioden mars 2019–februar 2020.

Språkpar	***Ordpar fra samisk***	***Ordpar til samisk***	***Søk (netto) i ett år***	***Prosentandel treff***	***Søk pr. taler (talere)***
nordsamisk↔norsk	44 906	36 775	1 436 218	86,5 %	72 (20 000)
nordsamisk↔finsk	13 198	16 055	442 971	84,3 %	261 (1700)
sørsamisk↔norsk	15 216	19 883	477 939	89,8 %	1593 (300)
enaresamisk↔finsk	29 353	30 630	322 246	89,7 %	806 (400)
skoltesamisk↔finsk	25 326	41 390	445 887	95,2 %	1486 (300)
skoltesamisk↔norsk	4003	6189	1111	82,4 %	-
skoltesamisk→engelsk	4782	0	1461	78,6 %	-
pitesamisk→svensk	5436	0	760	79,3 %	20 (30)

Kilde: Giellatekno UiT[16].

Feil! Fant ikke referansekilden. viser størrelsen på og bruken av de forskjellige NDS-ordbøkene[17]. Lulesamisk har ikke ordbok i NDS-format fordi Anders Kintels lulesamiske ordbok ikke har en struktur som gjør det mulig å bruke analyse med en språkmodell slik det blir gjort i NDS. Ordboka er likevel tilgjengelig som e-ordbok i ordboksappen *Julevbágo*, se kapittel 4.3.2.

Andre og tredje kolonne fra venstre gir antall ordpar for de aktuelle språkpara, neste kolonne viser antall søk i perioden. Ikke alle søka gir treff i ordboka, og fjerde kolonne angir prosentandelen av treff i ordboka. Alle ordbøkene gir forslag på ord når man begynner å skrive søkeordet, og dette er nok en medvirkende årsak til at også de minste ordbøkene har høy treffprosent. Når vi sammenlikner antall søk med stipulert antall samisktalende i henholdsvis Norge og Finland, kan vi beregne hvor mange ganger hver taler i gjennomsnitt har gjort et søk i ordboka i løpet av ett år (kolonnen helt til høyre).

Brukerloggene viser en markant forskjell mellom de samiske språka: Alle de mindre samiske språka bruker ordboka mye oftere enn nordsamisk. Spesielt stor er forskjellen mellom sør- og nordsamisk; talla for 2018 var henholdsvis 1911 og 53 oppslag per samiskspråklig (mellom disse språka og norsk). Dette er uttrykk for en generell trend: Mindre språksamfunn bruker språkressurser mer i forhold til antall språkbrukere enn større språksamfunn (jf. f.eks. Antonsen 2018:85), noe som

[16] For nettosøk har vi trukket fra IP-adresser som vi antar tilhører Googles søkeroboter. Overslag over antall samisktalere er fra http://www.ethnologue.com (Eberhard m.fl. 2020), for Finland fra https://www.samediggi.fi/saamelaiset-info/. Talla fra Sverige er ikke med i tabellen, men ifølge Ethnologue er det 4000, 1500 og 300 talere av henholdsvis nord-, lule- og sørsamisk.

[17] Ordbøkene er på disse internettadressene: https://sanit.oahpa.no https://baakoeh.oahpa.no https://saanih.oahpa.no https://saan.oahpa.no https://bahkogirrje.oahpa.no

speiler språksituasjonen med mindre støtte for språkbruken i dagliglivet, usikker skriftlig norm og en stor andel av brukere med samisk som andrespråk. Til sammenlikning blei *Bokmålsordboka* og *Nynorskordboka* på nett brukt gjennomsnittlig åtte ganger av hver norsktalende i 2017 (Antonsen 2018:2).

Eskonsipo (2020) viser at NDS for nordsamisk er i bruk mellom klokka 8 og 17, dvs. i arbeids- og skoletida, med en markant nedgang i bruk etter arbeidstid og i skole- og fellesferien. Hennes konklusjon er at i den grad ordboksbruk reflekterer faktisk produksjon av tekst, blir nordsamisk primært skrevet i arbeids- og skoletida.

4.3.2 Ordbøker uten bøyningsformer

Vi har tatt kontakt med en del aktører for å få nærmere informasjon og bruksstatistikk om enkelte allmenne samiske digitale ordbøker på internett, og også om noen ordboksapper. Ingen av ordbøkene i dette kapitlet kan analysere bøyde ord, slik som NDS, men *satni.org* gir bøyningsformer for orda i samlinga.

Nettstedet *sátni.org*, som drives av Divvun-gruppa, tilbyr flere ordbøker og samlinger av fagtermer i samme grensesnitt. I tillegg til ordboksmaterialet i NDS, tilbys, i samarbeid med Sámi Giellagáldu, lister med terminologi for ulike fagområder og på norsk, svensk, finsk og ulike samiske språk, som har til sammen 49 903 termer med oversettelser[18]. Søk det siste året går fram av **Feil! Fant ikke referansekilden.**.

Tabell 2. Antall søk i sátni.org i løpet av ett år (11.5. 2019–10.5.2020)[19].

Fra følgende språk	**norsk**	**finsk**	**nord-samisk**	**sør-samisk**	**lule-samisk**	**enare-samisk**	**skolte-samisk**
Antall søk på ett år	40 840	18 722	17 719	6419	6775	4417	3422
Antall søkeord	12 005	8398	7876	2753	1826	1899	1065

Kilde: Divvun-gruppa UiT.

Tabellen viser at søkertalla for satni.org er langt lavere enn for NDS. Dels er dette fordi satni.org retter seg mot profesjonelle brukere, og dels kan det være fordi samisk fagspråk er lite innarbeida. Som vist i en separat studie (Trosterud 2019) går rundt 60 % av ordboksoppslaga i NDS fra

[18] Termlistene er tilgjengelige på wikien https://satni.uit.no/termwiki.

[19] Ordboka fungerer slik at brukerne skriver søkeord uansett språk, og de får da opp alle tilgjengelige oversettelser. Det er derfor ikke mulig å skille mellom ulike språkpar.

minoritetsspråket (samisk) til majoritetsspråket (norsk, finsk). For termsamlingene er situasjonen motsatt, her går storparten av oppslaga fra majoritetsspråket.

Forlaget *Davvi Girji* har siden 2012, etter avtale med Sametinget, gjort to papirordbøker (Kåven mfl. 1995; SNSO 2000) tilgjengelige på internett. Men online-versjonene inneholder ikke forklaringer eller eksempler på bruk, slik som papirordbøkene gjør. Når man søker fra nordsamisk til norsk, får man informasjon om stadieveksling.

DinOrdbok[20] er ei online-ordbok som tilbyr oversettelse mellom 28 språkpar, blant disse også mellom norsk og nordsamisk, lulesamisk og sørsamisk. For disse tre språkpara brukes materiale med åpen kildekode, nemlig Giellatekno ordlistefiler og ordlister fra oversettingsprogramma på Apertium-plattformen, se kapittel 4.5.1.

Julevbágo[21] er ei lulesamisk–norsk–lulesamisk ordbok som bygger på Anders Kintels ordbok, som Sametinget i Norge har kjøpt bruksrettigheter til. Dette materialet finnes nedlastbart på internett, og Julevbágo tilgjengeliggjør dette som online-ordbok, og også via nedlastbar app. I tillegg inneholder Julevbágo noen termordlister fra Sámi Giellagáldu. Ordboka gir informasjon om stadieveksling, den inneholder eksempelsetninger, og man søker både i lista over grunnformer og i forklaringer og eksempler.

I Sverige finnes nettordbøkene *Sametingets ordböcker*. Den første versjonen med lule- og sørsamisk blei lansert 2008, nordsamisk kom med fra 2019. Ordbøkene bruker data fra det svenske Sametingets web-ordbok, Israelsson & Nejnes sydsamiske ordbok (2007), Nils Olof Sortelius' lulesamiske ordbok (2005) og Svonnis nordsamiske ordbok (2013). Det legges stadig flere ord til den lulesamisk ordboka.[22] Ordbøkene gir litt informasjon om ordbøyning, og for noen ord også kommentarer til betydningen, men det er ingen eksempelsetninger.

[20] https://www.dinordbok.no

[21] http://julev.no

[22] Informasjon om ordbøkene fått i e-post fra Anders Östergren Njajta 29.4.2020.

Tabell 3. Ordbøker online som vi har mottatt data for[23].

Utgiver	*Språkpar*	*Antall ordpar*	*Antall søk*	*Periode (ett år)*
Davvi Girji	nordsamisk–norsk	51 668	140 959	2019
DinOrdbok[24]	nordsamisk–norsk	36 000	58 000	2017
	lulesamisk–norsk	5000	43 000	2018
	sørsamisk–norsk	15 000	45 000	2018
Julevbágo	lulesamisk–norsk	18 500	569 000	29.2.19–28.2.20

Kilder: Davvi Girji v. Frank Rasmus og Jan Helge Soleng 3.3.2020, DinOrdbok v. Jon Atle Sandbakken 15.2.2020, Julevbágo v. Simon Paulsen 28.2.2020.

Feil! Fant ikke referansekilden. viser opplysninger om bruk av de samiske ordbøkene på internett som vi har mottatt data for. Vi mangler data for *Sametingets ordböcker*. Alle ordbøkene oversetter i begge retninger. Den tredje kolonnen viser antall ordpar, som er det samme antall i hver retning. Kolonne fire viser antall søk for ett år. Det varierer hvilket år vi har tall for. Vi kommenterer talla i kapittel 4.3.4.

4.3.3 Nedlastbare ordbøker

Noen av ordbøkene nevnt i forrige kapittel kan lastes ned som app til mobiltelefon og nettbrett eller datamaskin. Her kan vi bare angi hvor mange ganger ordbøkene er lasta ned, ikke hvor mye de er brukt.

Giellatekno tilbyr nedlastbare ordbøker, *Vuosttaš digisánit* for nordsamisk–norsk og *Voestes digibaakoeh* for sørsamisk–norsk (Antonsen mfl. 2009)[25], med samme innhold som NDS hadde i 2013, men bare med oversettelse fra samisk til norsk. De nedlastbare ordbøkene kan på nåværende tidspunkt ikke lastes ned med tilgang til grammatisk analyse og tekstsamlinger, slik som NDS online har. Derfor har disse ordbøkene ikke vært oppdatert på mange år. Det arbeides på sikt for heller å tilby NDS som nedlastbar ordbok med analyseprogram.

Sikku Media har gitt ut en nordsamisk–norsk ordboksapp *Samisk ordbok* som oversetter både til og fra nordsamisk, og som blei lansert 5.1.2018. Appen baserer seg på Giellateknos ordboksfiler fra dette tidspunktet og inneholder ca. 30 000 ordpar for hvert av språkpara. Ordboka gir ingen

[23] I disse talla er det ikke fjerna søk som Googles roboter har gjort, så de reelle brukertalla er nok en del mindre.
[24] Talla for ordpar i DinOrdbok er stipulert fordi de oppgitte talla på hjemmesiden også inneholder ordpar fra Apertium som er egennavn oversatt med samme egennavn, f.eks. *Hansen = Hansen*.
[25] Det finnes ikke tall for nedlastninger av Vuosttaš digisánit og Voestes digibaakoeh.

informasjon om bøyning av orda, ingen eksempelsetninger, men den gir litt tilleggsforklaringer til noen ordpar. Denne ordboksappen blei lasta ned 2080 ganger i perioden 5.1.2017–15.1.2020[26].

Sátnegirji er en nordsamisk–svensk ordboksapp som blei utgitt av Ravda lágádus i 2017, og som har samme innhold som Svonnis papirordbok (2013). Den tilbys nå dessuten online via Sametingets ordböcker. Også i appen kan man søke begge veger, og man får informasjon om stadieveksling, og litt ekstra forklaringer om bruken for enkelte ordpar, men det er ikke eksempelsetninger. Ordboksappen blei lasta ned 4274 ganger i perioden 2017–15.1.2020[27].

Fra mai 2014 har det også vært mulig å laste ned *Sametingets ordböcker* som app og bruke den offline. Den blei lasta ned 9159 ganger i åra 2014–2018. Innholdet i denne appen oppdateres automatisk når appen er online, slik at brukerne alltid får nyeste innhold av ordboka.

4.3.4 Drøfting av de ulike ordboksløsningene

Når vi sammenlikner brukstalla for ordbøker med bøyningsformer (NDS) med ordbøker uten slik tilleggsinformasjon, ser vi store forskjeller dem imellom. Ordbøkene mellom nordsamisk og norsk uten bøyingsformer har ca. 200 000 søk over ett år, mens NDS (**Feil! Fant ikke referansekilden.**) har halvannen million over samme tidsperiode, selv om Davvi Girjis ordbok inneholder betydelig flere ordpar enn NDS-ordboka. For sørsamisk–norsk er det en halv million søk for NDS og 15 000 for Din Ordbok. Dette viser at grammatikken som NDS-ordboka tilbyr, er populær. Det finnes ingen lulesamisk versjon av NDS, og dette gir utslag på talla i **Feil! Fant ikke referansekilden.**, hvor den lulesamiske Julevbágo har over en halv million søk, dvs. tilsvarende tallet for NDS for sørsamisk.

Ordbøkene til og fra andre samiske språk enn nordsamisk har mellom 10 000 og 30 000 oppslagsord, mens de største nordsamiske ordbøkene, Davvi Girjis ordbøker, har i underkant av 52 000 oppslagsord i hver retning. Ordforrådet i deres norsk–nordsamisk ordbok er basert på oversettelsene i nordsamisk–norsk ordbok, og store deler av det norske grunnordforrådet mangler dermed. For eksempel kan man ikke slå opp for å se hvordan det norske ordet *hverandre* skal oversettes til nordsamisk. For tospråklige allmennordbøker generelt blir 50 000 oppslagsord regna som et minimum for å dekke allmennspråklig tekst, store nordiske tospråklige ordbøker som de mellom svensk og finsk[28] inneholder for eksempel 110 000 oppslagsord hver veg. Dagens samiske ordbøker burde med andre ord ha vært større. Samtidig er det også behov for bedre ordboksartikler:

[26] E-post fra John Anders Sikku 15.1.2020.
[27] E-post fra Mikael Svonni 22.1.2020.
[28] Jf. Cantell mfl. 2004 og Karlsson (red.) 1982–87.

Mer informasjon om hvert oppslagsord og hvordan det brukes, og i tilfelle flere oversettelser trengs det også informasjon om i hvilke kontekster de ulike oversettelsene blir brukt. Dessverre har Davvi Girji valgt å ikke ta denne informasjonen fra papirordbøkene med inn i den elektroniske utgaven. NDS har slik informasjon, men ikke for alle oppslagsorda. Selv om NDS gir brukeren tilgang fra ordboka til setningseksempler i SIKOR (se kap. 4.4), vil det være behov for å belyse forskjellig bruk av ordet i selve ordboksartikkelen. Søk i SIKOR kan gi tusenvis av setningseksempler, men uten den sorteringa som brukeren trenger.

Til dette kommer så terminologi. For å illustrere behovet for terminologisk arbeid kan vi se på sjukepleie. UiT skal i nær framtid starte opp et utdanningstilbud i sjukepleie på nordsamisk. Sjukepleie har, som alle fag, sin egen terminologi. Det er for eksempel utarbeida en internasjonal terminologisk database for termer innafor sjukepleie[29]; den inneholder 4475 termer på norsk. Sámi Giellagáldus termdatabase har ingen egen kategori *Sjukepleie*, men kategorien *Medisin* inneholder 1535 norske termer med nordsamisk oversettelse. Bare 16 av termene i de to basene overlapper. Eksempelet illustrerer hvilke utfordringer samiske språk står overfor, på fagfelt etter fagfelt.

Enspråklige ordbøker finnes ikke for noe samisk språk, med unntak av ei papirordbok med synonymer for nordsamisk (Vest 2005). Denne mangelen utfordrer leseren til å spørre hvorfor det finnes enspråklige ordbøker i det hele tatt (som det gjør for f.eks. norsk og svensk). Svaret er at enspråklige ordbøker gjør brukerne i stand til å definere sitt eget språks begreper på sitt eget språk, uten å gå vegen via et annet språk. I tillegg vil enspråklige samiske ordbøker også samle hvert samiske språksamfunn over statsgrensene, i stedet for å være avgrensa til å få hver sin definisjon av det felles samiske ordforrådet, på hvert sitt majoritetsspråk.

Ordbøker mellom de samiske språka finnes nesten ikke. Et unntak er ei mindre ordbok mellom nordsamisk og kildinsamisk (Sammallahti & Xvorostuxina 1991). Det burde være mulig å slå opp ord på ett samisk språk for å finne det tilsvarende på et annet samisk språk uten å gå vegen om majoritetsspråka, og det at særlig nordsamisk blir styrka som fagspråk, øker behovet for slike ordbøker.

Gjennomgangen av samiske ordbøker viser at selv om ordboksmaterialet kan ha forskjellige kilder, har sametinga i stor grad finansiert trykking av papirordbøkene, og sametinga og UiT gjør ordbøkene tilgjengelige online. I tillegg til ordbøkene som er nevnt i forrige kapittel, kan det nevnes at Sametinget i Finland har kjøpt rettighetene til Pekka Sammallahtis nordsamisk–finsk ordbok (1993), og det arbeides ved UiT med å lage en online-utgave av denne.

[29] https://www.icn.ch/what-we-do/projects/ehealth/icnp-browser

Papirordbøker får færre og færre brukere, men elektroniske ordbøker gjør ordboksmaterialet mer tilgjengelig; det er alltid med i lomma via mobiltelefonen, og det er også lett å slå opp i det på datamaskinen. Dette gjør at et godt ordboksinnhold blir enda mer brukt enn tidligere. Loggdata for minoritetsspråksordbøker viser at jo mindre språksamfunnet er, desto større er behovet for ordbøker. En sentral del av arbeidet med minoritetsspråk bør derfor være å satse på bedre ordbøker.

4.4 Samiske tekstsamlinger på nett (SIKOR i Korp)

Ei god og stor tekstsamling (tekstkorpus) er et nyttig hjelpemiddel for arbeid med språknormering og utarbeiding av ordbøker og terminologi, og det er nyttig for oversettere, forskere og studenter. Størrelsen på et slikt tekstkorpus er avgjørende for hvorvidt det er mulig å utvikle språkteknologiske verktøy basert på statistikk og maskinlæring. I tillegg er det viktig å ha tekster som inneholder ordforrådet og typen språkbruk som passer for det verktøyet man vil utvikle. De samiske tekstkorpusa er små, sammenliknet med det norske NOWAC, som har 700 mill. ord, og med svensk og engelsk, hvor vi ikke snakker om millioner, men om milliarder av ord[30].

Det samiske tekstkorpuset SIKOR (Samisk Internasjonale korpus) inneholder digitale tekster for fem samiske språk. Korpuset er åpent tilgjengelig for korpussøk på internett via grensesnittet Korp. I underkant av halvparten av korpuset er tilgjengelig for nedlastning under en fri lisens.

Tabell 4. Antall ord for hvert språk og språkpar i SIKOR, tilgjengelig via søkegrensesnittet Korp.

Enspråklige tekster	***Antall ord i tekstene***	***Tospråklige tekster***	***Antall ord på hvert av språka***
nordsamisk	32 240 000	**norsk-nordsamisk**	3 480 000
lulesamisk	1 250 000	–	
sørsamisk	1 500 000	**norsk-sørsamisk**	198 000
enaresamisk	1 770 000	**finsk-enaresamisk**	85 000
skoltesamisk	213 000	**–**	

Kilde: http://gtweb.uit.no/korp versjon 06.11.2018.

SIKOR inneholder en stor del av all elektronisk tilgjengelig samisk tekst, innsamla dels direkte fra institusjoner som har produsert samisk tekst, og dels fra internett, se tabell 4. Tekstene er analysert med analyseprogramma som er beskrevet i kapittel 3.3, og dermed kan man f.eks. søke på

[30] Svensk Korp inneholder 13 milliarder ord, og Google Books 155 milliarder. Talla er fra 2018.

grunnform og få treff på setninger som inneholder alle bøyde former av ordet. Mer informasjon om den grammatiske analysen av korpuset er presentert i Antonsen & Trosterud (2017).

Brukere kan klikke seg direkte fra NDS-ordbøkene til Korp-grensesnittet og se hvordan ordet man er interessert i, brukes i setninger. Man kan også arbeide direkte i Korp. Det er en egen avdeling i SIKOR for tospråklige tekster, dvs. oversettelser fra majoritetsspråket til samisk, som er parallellisert på setningsnivå, dvs. at originalsetninga og den oversatte setninga presenteres ved siden av hverandre. Noen av disse er tilgjengelige i Korp, se tabell 4. Slik kan brukeren se hvordan termer og talemåter er oversatt til samisk av andre oversettere. Man kan også klikke seg til disse tekstene via NDS.

Tabell 5. Antall søk i enspråklige tekstsamlinger i SIKOR i perioden 28.1.–10.5.2020.

Språk	**nordsamisk**	**sørsamisk**	**skoltesamisk**	**enaresamisk**	**lulesamisk**
Antall søk	190 594	47 665	47 549	23 839	1195

Kilde: Giellatekno UiT.

Tabell 6. Antall søk i tospråklige tekstsamlinger i SIKOR i perioden 28.1.–10.5.2020.

Språkpar	**norsk–nordsamisk**	**norsk–sørsamisk**	**norsk–nord-/sørsamisk**	**finsk–enaresamisk**
Antall søk	83 931	18 942	23 405	21 021

Kilde: Giellatekno UiT.

Tabellene 5 og 6 viser statistikk for bruk av SIKOR. Til sammen mottok de ulike samiske tekstsamlingene 458 141 søk i løpet av denne perioden, eller i underkant av 17 000 søk pr. dag. Den alt overveiende delen av søka kommer via ordboksgrensesnittet. Loggdata skiller ikke mellom søk fra ordbok og søk direkte i SIKOR, men ved å sammenligne talla for sør- og lulesamisk ser vi hva som skjer. Korpusa for disse språka er omtrent like store, og lulesamisk har flere talere enn sørsamisk. Likevel får det lulesamiske korpuset bare 1195 søk i perioden, noe som utgjør 2,5 % av det sørsamisk får (47 665). Den store forskjellen er at det for sørsamisk finnes ei ordbok som lenker til det sørsamiske korpuset, mens det ikke gjør det for lulesamisk. SIKOR blei opprinnelig laga for språkforskere, og muligheten for å slå opp i korpuset via ordbok kom senere. Statistikken viser likevel klart at de samiske språksamfunna er ei langt viktigere brukergruppe for samisk språkteknologi enn det språkforskerne er.

4.5 Oversettingsverktøy

Oversettere for majoritetsspråka har siden 1980-tallet brukt både maskinoversetting og såkalt *oversettingsminne*, dvs. tospråklige tekstsamlinger til hjelp i oversettingsarbeidet. De siste 15 åra

har maskinoversetting blitt tilgjengelig for mange språkpar via selskap som Google og Microsoft, og datateknologien har dermed fått ei mer sentral rolle i flerspråklige sammenhenger. Både oversettingsminne og maskinoversetting har de siste åra blitt tilgjengelige også for samiske språk.

4.5.1 Maskinoversetting

Maskinoversetting kan brukes for å forstå en tekst, og for leseren er det da viktigere at alle orda blir riktig oversatt enn at oversettelsen har et godt språk. Maskinoversetting kan også brukes til å produsere tekster, som for begrensede tekstdomener, som menyer, værmeldinger og bruksanvisninger, hvor det er viktig å bruke riktig term, og hvor språket har liten variasjon. Dessuten kan maskinoversetting brukes til såkalt *posteditering*, det vil si at oversettelsen er en kladd som en menneskelig oversetter forbedrer. Da er hensikten å spare tid, og kanskje også å hjelpe oversetteren å være konsistent i ordvalg, hvis det er en type tekst som krever dette. Et slikt program er ikke tilpassa oversetting av skjønnlitterære tekster, hvor det er viktigere å variere ordbruken og å finne formuleringer som er typiske for språket man oversetter til. Skjønnlitteratur har også større ordforråd enn andre tekster og krever et mye mer utbygd oversettingsprogram enn andre typer tekster.

UiT arbeider med oversettingsprogram fra nordsamisk til norsk, lulesamisk, sørsamisk og enaresamisk. Programma bruker regelbasert teknologi og *Apertium*-plattformen. Plattformen egner seg spesielt for språk som ligger nær hverandre grammatisk, og brukes ikke minst for oversetting fra bokmål til nynorsk (bl.a. av Nynorsk pressekontor, Wikipedia og skoleelever) og fra spansk til katalansk, der den også brukes til oversetting av artikler i dagsavisa *La Voz de Galicia*. I tillegg brukes Apertium mellom enkelte språkpar på Wikipedia.

Tekst på kildespråket legges inn i oversettingsprogramma og analyseres med en språkmodell (se kapittel 3.3). Grunnorda byttes ut med målspråkets ord, og så bygger programmet setninga på målspråket ved hjelp av manuelt skrevne regler. Til sammenlikning er *Google Translate* basert på maskinlæring, noe som krever atskillig større samlinger med grunntekst og oversettelse enn det som finnes for samiske språk.

Tabell 7 viser hvor mange ordpar som finnes i ordlista i oversettingsprogramma. Det er atskillig flere ordpar for nordsamisk–norsk enn det er mellom de samiske språka. Det er arbeida mest med programmet for nordsamisk–norsk, men den viktigste grunnen er at svært mange ord har samme oppbygning på de forskjellige samiske språka, slik at programmet klarer å bygge dem ved hjelp av ordsammensetning eller ordavledning, eller en kombinasjon av disse. Fra nordsamisk til norsk er man i mye større grad avhengig av ord-til-ord-oversetting. F.eks. bruker alle de samiske

språka i Tabell 7 avledninger av ordet for å skrive *čállit → čálli, čäälliđ → čällee, tjaeledh → tjaelije, tjállet → tjálle*, henholdsvis på nordsamisk, enaresamisk, sørsamisk og lulesamisk, for å uttrykke det som på norsk oversettes med *forfatter* eller *sekretær*.

Tabell 7. Antallet ordpar i hvert av Apertiums maskinoversettingsprogram[31].

Språk	**nordsamisk–norsk**	**nordsamisk–lulesamisk**	**nordsamisk–sørsamisk**	**nordsamisk–enaresamisk**
Antall ordpar	42 216	11 220	6995	8743

Kilde: https://github.com/apertium 31.01.2020.

Ordlistene i oversettingsprogramma kan bli lasta ned fra internett, og det er også gjort for DinOrdbok, se kapittel 4.3.2. Men det er viktig å være oppmerksom på at mange av ordpara bare er ment brukt i helt spesielle sammenhenger. I tillegg er det omtrent 46 000 ordpar med egennavn for hvert språkpar. Både for å få til en full analyse av setningene, og for å kunne legge til riktig kasusending, må man i ordlista også ha navn som har samme form på kildespråket og målspråket, så som *Hansen = Hansen*.

Maskinoversetting for samisk er tilgjengelig via flere kanaler. Nettsidene[32] tilbyr tre muligheter: kopiere/skrive inn tekst i ei dialogrute, oversette et helt dokument eller oversette ei nettside. I tillegg er det mulig å installere oversettingsprogramma i oversettingsverktøyet *OmegaT*, og det er i prinsippet mulig å laste ned og installere programmet til eget bruk fra *github* (det siste krever godt kjennskap til Giellatekno-infrastrukturen).

I løpet av januar 2020 leverte nettsida for oversettelse mellom nordsamisk og norsk 28 167 oversettelser av tekst som brukeren hadde kopiert inn i dialogruta. I samme periode oversatte programmet også 1930 nettsider. Nesten alle de oversatte nettsidene (98 %) var Samisk høgskole sine sider. De to andre sidene som var representert, var det samiske vitenskapelige tidsskriftet *Sámi dieđálaš áigečála* og *Senter for samisk i opplæringa*. At Samisk høgskole dominerer, er ikke uventet, siden dette er det eneste nettstedet som har dirigert brukerne til en norsk maskinoversatt versjon via en ferdig snarveg. For å oversette andre sider må brukeren selv lime inn lenka til nettsida, noe mange sikkert ikke er klar over.

Til tross for at tekster til Samisk høgskoles nettsider blei oversatt i gjennomsnitt 62 ganger hver dag i januar 2020, har høgskolen nå fjerna direktelinken. Begrunnelsen er at de synes den

[31] Ordpar med egennavn kommer i tillegg til talla i tabellen.
[32] http://jorgal.uit.no/ (nordsamisk → norsk) og http://gtweb.uit.no/mt/ (mellom samiske språk)

språklige kvaliteten på oversettelsene ikke var bra nok, og de var bekymret for at brukere av nettsida skulle tro at dette var høgskolens manuelle oversettelse.

I enkelte diskusjonsgrupper om samiske forhold på Facebook har det flere ganger vært diskusjon om bruk av samisk språk. De som ikke behersker språket, har følt seg utestengt fra diskusjonen hvis noen har skrevet en kommentar på samisk. Nå er det noen grupper som viser til oversettingsverktøyet for dem som trenger det, og uten å ha noen statistikk på dette virker det som at det nå skrives mer på nordsamisk på Facebook enn tidligere. Dette er et eksempel på at maskinoversetting styrker samisktalendes muligheter til å bruke samisk også i offentlige sammenhenger. Et annet eksempel er at den ene forfatteren av denne artikkelen kunne skrive sin PhD-avhandling på nordsamisk fordi deler av vurderingskomiteen kunne bruke maskinoversetting til hjelp i lesinga.

Oversettingsprogramma fra nordsamisk til andre samiske språk kan gjøre viktige nordsamiske tekster tilgjengelige for andre samiskspråklige, og dermed styrke det faglige fellesskapet mellom de ulike samiske språksamfunna. Tekst skrevet på nordsamisk har ofte ei problemstilling og en innfallsvinkel som er relevant for andre samiske språksamfunn, og ved å kunne oversette teksten til sitt eget samiske språk vil talere av andre samiske språk kunne lese den. Programma kan også brukes motsatt veg, til nordsamisk, selv om programma i dag bare tilbyr en enkel ord-for-ord-oversettelse denne vegen. Men ei slik oversetting til nordsamisk gjør det mulig for studenter å skrive oppgaver og artikler på lule-, sør- og enaresamisk, og sensorer og kolleger som bare kan nordsamisk, kan med støtte av en slik maskinoversatt tekst være i stand til å lese tekstene deres.

Oversettingsprogramma som er tilgjengelige i dag, er imidlertid ikke gode nok, de må videreutvikles. Det vil i praksis si at programma må tilpasses forskjellige sjangre og domener, og dette gjøres best i samarbeid med oversettere og miljøer som produserer tekst som skal oversettes.

4.5.2 Oversettingsminne og datastøtta oversetting (CAT)

Som en del av arbeidet med SIKOR har Giellatekno og Divvun også samla inn tekstsamlinger for ulike språkpar mellom samiske og andre språk og parallellisert dem setningspar for setningspar. Den største tekstsamlinga blei samla inn med finansiering av det som da var Forbruker- og administrasjonsdepartementet (jf. Trosterud & Eskonsipo 2013). Seinere er mer tekst og flere språkpar lagt til. Disse parallelle tekstsamlingene er tilgjengelige via NDS-ordbøkene, og for sammenliknende språkforsking via grensesnittet Korp. I tillegg er de åpne parallelle tekstsamlingene nedlastbare som oversettingsminne, jf. **Feil! Fant ikke referansekilden.**.

Tabell 8. Tilgjengelige parallelle tekstsamlinger og ordlister for bruk i datastøtta oversetting (CAT).

Språkpar	***Setningspar***	***Antall ord (språk 1)***	***Antall ordpar i ordlista***
norsk–nordsamisk	216 456	3 391 942	68 151
norsk–sørsamisk	7166	72 741	14 940
norsk–lulesamisk	1643	16 272	11 893
finsk–nordsamisk	109 852	1 162 221	9647
finsk–enaresamisk	34 351	366 540	30 211
finsk–skoltesamisk	6481	70 087	42 150
nordsamisk–sørsamisk	23 746	256 070	553
nordsamisk–lulesamisk	18 244	194 907	412
nordsamisk–enaresamisk	13 514	150 516	6193

Kilde: https://giellalt.uit.no/tm/TranslationMemory.html 10.02.2020.

Tekstsamlingene er alle lagra i et format som gjør at de kan brukes av all programvare for datastøtta oversetting. Også NDS-ordbøkene er nedlastbare som ordlister i et format som passer programma for datastøtta oversetting. Generelt sett er slike program svært utbredt blant oversettere, men så langt har samiske oversettere utgjort et unntak. I løpet av det siste året ser det likevel ut til at samiske oversettere vil ta i bruk program for datastøtta oversetting. Oversetting av tekst til samisk utgjør en flaskehals i svært mange sammenhenger, særlig når det gjelder skolebøker og sakprosa, så mer effektive rutiner for oversetting vil ha stor innflytelse på språksamfunnet. Bruk av oversettingsminne og automatisk oppslag av fagtermer vil også gjøre det lettere for oversetterne å få en mer konsistent terminologi.

Mange program for datastøtta oversetting tilbyr ordbøker som en av flere typer ressurser til hjelp for oversetteren. Samiske oversettere vil dermed trenge ordbøker for språka de oversetter fra. Ett ordboksformat som er mye brukt, er *StarDict*. For finsk er det ei enspråklig ordbok tilgjengelig i StarDict-format[33], dette gjelder dessverre verken for norsk eller svensk.

4.6 Syntetisk tale og talegjenkjenning

Divvun publiserte i 2015 et program for nordsamisk syntetisk tale. Det er i bruk i en del sammenhenger, f.eks. på nettsidene til Troms og Finnmark fylkeskommune og flere kommuner

[33] jf. https://sites.google.com/site/gtonguedict/home/stardict-dictionaries

innafor det samiske forvaltningsområdet. Dessuten er det i bruk som hjelpemiddel for folk som har vansker med å lese vanlig skrift, f.eks. på forberedelsesdelen til grunnskoleeksamen i nordsamisk som førstespråk[34] og til høytlesing av tekster som alternativ til punktskrift (Statped: Samisk punktskrift). Statped (Statlig pedagogisk tjeneste) arbeider med å få syntetisk tale for nordsamisk integrert i forskjellig programvare[35]. Ut over det er syntetisk tale for samisk ikke så mye i bruk. En mulig årsak til det kan være at teknologien er for tidlig ute: Syntetisk tale for norsk spiller ei relativt marginal rolle i sammenhenger der vi også finner samisk, og så lenge det er tilfelle, kan det hende at etterspørselen etter syntetisk tale for samisk forblir lav. På den andre sida har UiT planer om å ta syntetisk tale i bruk i ordbøker og språklæringsprogram.

Når det gjelder talegjenkjenning, finnes det fremdeles ikke fungerende program for noe samisk språk. Aalto-universitetet i Helsingfors har arbeida med samisk talegjenkjenning, og Divvun-gruppa har i 2020 fått ekstra finansiering for å arbeide med dette. Det er allerede interesse for å få talegjenkjenning, og denne interessen vil sannsynligvis øke når taleteknologi sprer seg til nye bruksområder for norsk.

4.7 Bruk av samisk språkteknologi på andre områder

Å ha ei analysert tekstsamling tilgjengelig for Korp som tekstanalyse har vært nyttig for samisk språkforsking. Tre av de fire siste artiklene om nordsamisk syntaks i *Sámi dieđálaš áigečála* (2014–2019) bruker f.eks. data fra SIKOR. UiT har også gjort andre ressurser tilgjengelige for språkforskere. Det inkluderer lister, laget ved hjelp av analyseverktøy, over hvor ofte ord, inkludert ulike bøyingsformer, forekommer i tekst (frekvensordlister). Man kan også bruke online-verktøy for grammatisk analyse. Maskinoversettingsprogrammet fra nordsamisk til norsk har også styrket vitenskapelig publisering på samisk: Nå vet samiskspråklige språkforskere at også norskspråklige kolleger kan lese publikasjonene deres.

Språkteknologi brukes også i språklæringsverktøy. Teknologien kan brukes som hjelpemiddel for å finne eksempelsetninger fra en tekstsamling og å gi bøyningsformer av ord, noe som også gjøres i NDS-ordbøkene, og dermed gir disse ordbøkene mulighet for mer språklæring enn det som er vanlig i ei ordbok. Man kan også analysere tekster online, eller det som brukeren skriver, og slik gi tilbakemelding til brukeren. UiT har utvikla program for å lære grammatikk, som *Oahpa* for nord- og sørsamisk og *Konteaksta* for nordsamisk, der språkteknologien blir brukt på denne måten. Bruken av språklæringsverktøy er et for stort tema til å behandles i denne artikkelen.

[34] Informasjon i e-post fra Astrid Eggen, Udir 12.05.2020.

[35] Samtale med Lone Nergård Boine 13.5.2020.

4.8 Utviding til andre språk

All språkteknologi utvikla ved Institutt for språk og kultur ved UiT er tilgjengelig som åpen kildekode. Dette har flere konsekvenser. Flere av de andre språkteknologiske initiativa for samisk, som e-ordbøkene fra DinOrdbok, bygger på åpne ressurser fra Giellatekno og OPUS[36]. I og med at kildekoden er åpen og språkuavhengig, er det også mulig å utvide den til flere språk. Dette har også blitt populært. I tillegg til de tre samiske språka som blir snakka i Norge, har Giella-infrastrukturen ført til stavekontrollprogram for 19 andre språk[37]. Ordboksportalen Neahttadigisánit, som opprinnelig blei laga for nordsamisk–norsk og sørsamisk–norsk (se Johnson mfl. 2013), inneholder nå 111 språkpar fordelt på 18 forskjellige ordboksportaler. Dette viser at det er stort behov og interesse for språkteknologiske løsninger for minoritetsspråk med omfattende bøyingsmønster og ordavledning, og at det ikke er så mange steder det er mulig å få tilgang til en infrastruktur som kan gi språksamfunna de verktøya de vil ha. Slik sett har satsinga på samisk språkteknologi fått resultater langt ut over det som var den opprinnelige intensjonen med satsinga.

5 Drøfting og konklusjon

Målbevisst arbeid med å integrere samiske bokstaver i internasjonale standarder gjorde det mulig å få samisk på internett allerede på 1990-tallet. Ordretteprogram og mobiltastatur blei tilgjengelige fra og med 2007 og framover. De har blitt lasta ned av store deler av de samiske språksamfunna og har dermed blitt en del av hverdagen for de aller fleste som skriver samisk. Etter forbedringa av samisk tekst på *Facebook* å dømme har innføringa av samiske tastatur på mobiltelefonen gjort det lettere å skrive samisk ikke bare på Facebook, men også i tekstmeldinger og annen tekstbruk på mobil.

Brukerloggene for NDS viser at ordbøker med bøyning er viktige for minoritetsspråk med omfattende bøyingsmønster. Rundt ett av ti ordbokssøk blir fulgt opp med et søk i autentisk samisk tekst. Når søka blir korrigert for størrelsen på språksamfunnet, viser ordboksloggene langt høyere tall for de mindre samiske språka enn for nordsamisk. Jo mindre språket er, desto mer har man behov for ordbøker og liknende ressurser. Det er viktig at de ulike programma er integrert i hverandre, noe vi ser av bruken av de elektroniske tekstsamlingene: De tilgjengelige tekstsamlingene for sør- og lulesamisk er omtrent like store. For begge språka finnes det digitale

[36] http://opus.nlpl.eu

[37] http://divvun.org/proofing/proofing.html

ordbøker, men bare for sørsamisk er det mulig å søke i tekstsamlinga via ordboksgrensesnittet. Brukerloggen viser da også at den sørsamiske tekstsamlinga er 40 ganger mer i bruk enn den lulesamiske. Slik ser vi hvordan tilgjengeligheten av språkteknologiske ressurser styrer folks adferd. Samtidig er realiteten at selv om det er de minste samiske språka som trenger verktøya mest, er det nordsamisk som har flest språkteknologiske verktøy, fordi både den best utbygde språkmodellen og de største tekstsamlingene er nordsamiske.

Et annet eksempel på slik endra atferd er maskinoversetting. Da maskinoversetting fra nordsamisk til norsk blei tilgjengelig, forsvant også store deler av diskusjonen om det skulle være lov til å skrive på samisk i nettdiskusjoner eller ikke. Maskinoversetting kan også bidra til at de samiske språka kan styrke hverandre, heller enn å kommunisere via majoritetsspråket, noe vi allerede ser eksempler på i akademisk sammenheng.

I et globalt perspektiv står de samiske språka i en langt sterkere situasjon enn språk med like mange talere og like stor grammatisk kompleksitet. Det er flere grunner til dette: De samiske språka er i bruk som skriftspråk, både i skole og administrasjon, og initiativ og finansiering til å lage et ordretteprogram for nordsamisk kom da også fra Sametinget i 2004. Samtidig var det i Tromsø et lingvistisk miljø som allerede var i gang med å lage språkmodeller for nordsamisk, inspirert av språkteknologiske modeller for finsk. Endelig fantes det en vilje i den norske statsadministrasjonen og på UiT til å finansiere et forskings- og utviklingsmiljø på omtrent ti personer. Den sentrale aktøren for utvikling av samisk språkteknologi har med andre ord vært UiT, med de konkrete behova for språkstøtteprogram i de samiske språksamfunna som pådriver.

Det er viktig å understreke at det bak alle de samiske språkmodellene ligger en hundreårig tradisjon med ordboksarbeid, der både privatpersoner og akademisk ansatte har gjort et stort arbeid for å nedtegne ordforrådet for alle de samiske språka. Uten dette arbeidet ville det språkteknologiske grunnarbeidet heller ikke ha vært mulig. Språkteknologisk utvikling for samiske og andre språk uten kommersielt potensial er avhengig av ikke-kommersielle aktører. I Norge har dette vært mulig, og resultatet har blitt fungerende språkteknologiske program også for samiske og andre språk utafor Norge.

De siste tiåra har dermed sett ei utvikling av språkteknologiske verktøy til støtte for den samiske skriveprosessen, først og fremst ordretteprogram, men også interaktive ordbøker og tilgang til autentiske språkeksempel via korpusgrensesnittet Korp. Dette har gjort skriveprosessen lettere og mer effektiv. De høye talla for både nedlastning og bruk viser at verktøya har en sentral plass i hverdagen til det store flertallet av dem som skriver på samisk. Spesielt andrespråkstalere gir uttrykk for at de ikke ville ha kunnet skrive på samisk uten tilgang på slike verktøy. Forskning på aktuell språkbruk faller utafor rammene til denne artikkelen, men på grunnlag av statistikker over

både nedlastning og bruk av ulike språkteknologiske program, og den språkbruken vi kan observere, konkluderer vi med at samisk språkteknologi har endra folks språklige atferd, og dermed de samiske språksamfunna.

Hvis vi ser på framtidsperspektiva for samisk språkteknologi, peker disse punkta seg ut som naturlige satsingsområder:

- Store deler av den samiske skriftkulturen, framfor alt skjønnlitteraturen, er fremdeles ikke tilgjengelig for språkteknologisk forskning og utvikling, men det bør den bli. Samiske språk kan ikke henge med i språkteknologisk utvikling framover uten store tekstsamlinger.
- Det trengs målretta arbeid med å forbedre ordboksressursene for samisk, særlig enspråklige samiske ordbøker og fagspesifikk terminologi, men også ordbøker de samiske språka imellom. Alle de nordiske nasjonalspråka har det til felles at de trenger gode leksikografiske ressurser, samtidig som de nordiske språksamfunna er for små til å finansiere dette arbeidet på kommersiell basis. For å løse dette har samtlige nordiske land oppretta egne leksikografiske institusjoner. For norsk er den lagt til Universitetet i Bergen. Selv med færre talere har de samiske språka det samme behovet for å strukturere og utvikle ordforrådet sitt. Det trengs med andre ord et «Samisk leksikografisk institutt».
- Språkmodellene som alt språkteknologisk arbeid hviler på, må bli bedre. Alle samiske språk bør opp på det samme nivået som nordsamisk.
- Det er mulig å lage langt bedre og mer intelligente skrivestøtteprogram enn det de samiske språka rår over i dag. Slike program vil styrke den samiske skriftkulturen.
- Samisk taleteknologi er med dagens syntetiske tale for nordsamisk så vidt kommet i gang. Neste steg her er syntetisk tale for de andre samiske språka, og system for å kjenne igjen samisk tale.
- Arbeidet som blir gjort for samiske språk, er relevant for alle morfologisk komplekse språk uten kommersielt potensial, dvs. for brorparten av språka i verden. Også i dette perspektivet er samisk språkteknologi et viktig satsingsområde.

Oppsummeringsvis er de viktigste erfaringene etter tjue år med samisk språkteknologi at gode språkteknologiske løsninger er viktige for alle språksamfunn, og viktigere jo færre talere et språk har. Sentrale verktøy som ordretteprogram og maskinoversetting gjør det lettere å bruke samisk; for mange språkbrukere ville det ha vært umulig å uttrykke seg skriftlig på samisk uten disse programma. Den målretta satsinga på samisk språkteknologi har ikke bare styrka den samiske skriftkulturen, den har også gitt opphav til en språkteknologisk infrastruktur som blir tatt i bruk for flere og flere språk, særlig språk med typologiske trekk felles med samisk. Samtidig er det nok av

utfordringer å ta fatt på. Bare et kontinuerlig og målretta arbeid kan videreutvikle og forbedre dataprogram for samiske språk og andre minoritetsspråkssamfunn, og dermed gi dem tilgang til den samme hjelpa datateknologien kan tilby majoritetsspråksamfunna.

Referanser

Antonsen, Lene (2013). Čállinmeattáhusaid guorran. [English summary: Tracking misspellings.] – *Sámi dieđalaš áigečála* 2/2013. s. 7–32. URL: https://site.uit.no/aigecala/sda-2-2013-lene-antonsen/

Antonsen, Lene (2018). *Sámegielaid modelleren – huksen ja heiveheapmi duohta giellamáilbmái* [English summary: Modeling Saami languages. Construction and adaptation to real-world linguistic issues]. PhD-avhandling, UiT Norges arktiske universitet. URL: https://munin.uit.no/handle/10037/12884

Antonsen, Lene, Gerstenberger, Ciprian, Moshagen, Sjur N. & Trosterud, Trond (2009). Ei intelligent elektronisk ordbok for samisk. – *LexicoNordica* Volum 16. Oslo: Nordisk forening for leksikografi. s. 271–283. URL: https://tidsskrift.dk/lexn/article/view/18479

Antonsen, Lene & Trosterud, Trond (2010). Manne dihtor galgá máhttit sámegiela? [English summary: Why the computer should know its Sami grammar?] – *Sámi dieđalaš áigečála* 1/2010. s. 3–28. URL: https://site.uit.no/aigecala/sda-1-2010-antonsen-ja-trosterud/

Antonsen, Lene & Trosterud, Trond (2017). Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. – *Norsk lingvistisk tidsskrift* Volum 35:1. s. 153–185. URL: http://ojs.novus.no/index.php/NLT/article/view/1416

Cantell, Ilse, Martola, Nina, Romppanen, Birgitta & Sundström, Mats-Peter (red.) (2004). *Suomi-Ruotsi-suursanakirja / Stora finsk-svenska ordboken*. 3., tarkistettu ja päivitetty painos, 2. laitos. Helsinki: WSOY. ISBN 951-0-24714-6

Eberhard, David M., Simons, Gary F. & Fennig, Charles D. (red.) (2020). *Ethnologue: Languages of the World*. Twenty-third edition. Dallas, Texas: SIL International. URL: http://www.ethnologue.com

Eskonsipo, Berit Merete Nystad (2020). *Sátnegirjegeavaheami čalmmustahttin neahttasátnegirjji loggafiilla analysa bokte*. Master i samisk språkvitenskap. Institutt for språk og kultur, UiT Norges arktiske universitet. URL: https://hdl.handle.net/10037/18505

Israelsson, Per-Martin & Nejne, Sakka (2008). *Svensk–sydsamisk, sydsamisk–svensk ordbok och ortnamn = Daaroen-åarjelsaemien, åarjelsaemien-daaroen baakoegärja jih sijjienommh*. Kiruna: Saemiedigkie.

Johnson, Ryan, Antonsen, Lene & Trosterud, Trond (2013). Using finite state transducers for making efficient reading comprehension dictionaries. *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NoDaLiDa 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16. s. 59–71. URL: https://www.aclweb.org/anthology/W13-5610/

Karlsson, Göran (red.) (1982–87). *Iso ruotsalais-suomalainen sanakirja. Stora svensk–finska ordboken (I–III)*. Suomalaisen Kirjallisuuden Seuran toimituksia 358. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kåven, Jernsletten, Nordal, Eira & Solbakk (1995). *Sámi–dáru sátnegirji = Samisk–norsk ordbok*. Kárášjohka: Davvi Girji.

Melhus, Marita & Broderstad, Ann Ragnhild (2020). *Folkehelseundersøkelsen i Troms og Finnmark. Tilleggsrapport om samisk og kvensk/norskfinsk befolkning*. Tromsø: Senter for samisk helseforskning, UiT Norges arktiske universitet.

Morottaja, Petter, Olthuis, Marja-Liisa, Trosterud, Trond & Antonsen, Lene (2018). Anarâškielâ tivvoomohjelm – Kielâ- já ortografiafeeilâi kuorrâm tivvoomohjelmáin. *Dutkansearvvi dieđalaš áigečála*, Volum 2018 (2). s. 63–84. URL: http://dutkansearvi.fi/volume-2018-issue-2-ps/

Nielsen, Konrad (1932–1962). *Lappisk ordbok – grunnet på dialektene i Polmak, Karasjok og Kautokeino*. 1–5. Oslo: Aschehoug.

Rueter, J. & Hämäläinen, M. (2020). FST Morphology for the Endangered Skolt Sami Language. —— *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*. European Language Resources Association (ELRA). s. 250–257. URL: https://arxiv.org/abs/2004.04803

Sammallahti, Pekka & Mosnikoff, Jouni (1991). *Suomi–koltansaame-sanakirja = Lää'dd-sää'm sää'nnǩe'rjj*. Utsjok: Girjegiisá.

Sammallahti, Pekka & Xvorostuxina, A. (1991). *Unna sámi-cāмь cāмь-sámi sátnegirjjáš*. Ohcejohka.

Sammallahti, Pekka (1993). *Sámi-suoma-sámi sátnegirji = Saamelais-suomalais-saamelainen sanakirja*. Ohcejohka: Girjigiisá.

Sjaggo, Ann-Charlotte & Trosterud, Trond (2015). Om pitesamiskt språk. Bjørg Evjen & Marit Myrvoll (red): *Från kust til kyst = Áhpegáttest áhpegáddáj*. s. 223–231. Tromsø: Orkana Forlag.

Statped: *Samisk punktskrift*. URL: https://www.statped.no/samisk-spesialpedagogisk-stotte/fagartikler/samisk-punkskrift/

SNSO = *Stor norsk–samisk ordbok = Dáru-sámi sátnegirji*. 2000. Kárášjohka: Davvi Girji.

Svonni, Mikael (2013). *Sátnegirji : davvisámegiela-ruotagiela, ruotagiela-davvisámegiela = Ordbok : nordsamisk-svensk, svensk-nordsamisk*. ČálliidLágádus.

Trosterud, Trond & Nystad Eskonsipo, Berit Merete (2012). A North Sami translator's mailing list seen as a key to minority language lexicography. *Euralex 2012 Proceedings*. s. 250–256. URL: https://euralex.org/publications/a-north-sami-translators-mailing-list-seen-as-a-key-to-minority-language-lexicography/

Trosterud, Trond (2012). A restricted freedom of choice: Linguistic diversity in the digital landscape. *Nordlyd* Vol 39, No 2. s. 89–104. URL: https://doi.org/10.7557/12.2474

Trosterud, Trond (2019). Kva bruker vi minoritetsspråksordbøker til? Ein studie av brukarloggane for tolv tospråklege ordbøker. *LexicoNordica* Volum 26. URL: https://tidsskrift.dk/lexn/article/view/117526

Valtonen, Taarna & Olthuis, Marja-Liisa (2018). *Suomi-Inarinsaame sanakirja*. Inari: Sámitigge.

Vest, Jovnna-Ánde (2005). *Sámi synonymat*. Anár: Sámediggi.

Wiechetek, Linda, Moshagen, Sjur N., Gaup, Børre & Omma, Thomas (2019). Many shades of grammar checking – Launching a constraint grammar tool for North Sámi. – *Linköping*

Electronic Conference Proceedings 2019 (168). s. 35–38. URL: https://visl.sdu.dk/pdf/CG-workshop2019_paper_1.pdf

Wilbur, Josh (red.) (2016). *Pitesamisk ordbok, samt stavningsregler*. Freiburg: Department of Scandinavian studies.