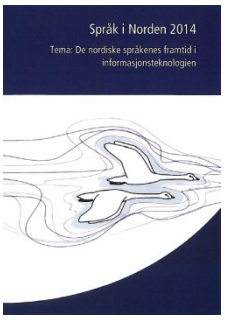
Sprog i Norden

Titel: Status for samisk språkteknologi

Forfatter: Sjur Nørstebø Moshagen

Kilde: Sprog i Norden, 2014, s. 95-109

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed

Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Status for samisk språkteknologi

Sjur Nørstebø Moshagen

Heilt frå utviklinga av den aller fyrste språkteknologien – det skrivne språket – og fram til i dag har språkteknologien vore både ei hjelp og ei grense. Hjelp for dei som har hatt han, og eit hinder for deltaking på like vilkår for dei som ikkje har hatt han, og ofte som eit middel for språkundertrykking overfor dei som ikkje har hatt han. Artikkelen gjev ei kort innføring i historia til samiske språk i eit språkteknologisk ljos, og korleis utviklinga av språkteknologi har vore med på å forma relasjonen til andre språk. Deretter gjev artikkelen ei oversikt over statusen for samisk språkteknologi i 2014, kva som er dei viktigaste ideane bakom arbeidet som er gjort, og kva som bør gjerast framover.

Bakgrunn

Hovudtanken bak arbeidet med samisk språkteknologi er at språka skal overleva og ha ei framtid som bruksspråk for heile det samiske samfunnet. Eit språk som ikkje blir brukt, er i praksis eit daudt språk. I og med at samfunnet er så gjennomdigitalisert som det er, betyr det at for å kunna brukast må dei samiske språka vera tilgjengelege og brukande i alle digitale samanhengar.

Språkteknologi og samfunn

Føresetnadene for å nytta språk, og samanhengane dei blir nytta i, har vorte dramatisk endra over dei siste fem hundre åra, og aller mest dei siste hundre. Dette heng i hop med den språkteknologiske utviklinga i vid forstand, og dei endringane som ho har ført med seg for samfunnet.

Kort skissert kan utviklinga summerast opp i desse punkta:

- utviklinga av skriftspråk og alfabet, med ulik teknologi (kiler og leire, kniv og tre, penn og papir)

- utviklinga av boktrykkjarkunsten

- utviklinga av massemedium og innføringa av massekommunikasjon
- utviklinga av datateknologi

Før noko av dette vart utvikla, var bruken av språk berre avhengig av menneska sjølve. Talen var den einaste modusen (eg ber teiknspråklege om orsaking for at eg forenklar litt), og overføringa av eit språk frå ein generasjon til neste gjekk naturleg og sjølvsagt.

Men med innføringa av språkteknologi endra dette seg. Kvart steg i den teknologiske utviklinga har ført til nye barrierar i bruken av språka, som oftast kopla til dei samfunnsmessige endringane som fylgde med den teknologiske utviklinga.

Med innføringa av skriftspråk kom det fyrste skiljet, mellom dei som har eit skriftspråk og dei som ikkje har. Dei fyrste samiske tekstene er frå 1600-talet, og med ordlister og bibelomsetjingar utover 1700- og 1800-talet. Ein eigentleg skriftspråkstradisjon kan ein ikkje prata om før kring 1900, og sjølv då er han svært avgrensa. Enno i dag er det samiske språk som ikkje har ei etablert skriftnorm, til liks med nesten fem tusen andre språk[1].

Innføringa av skriftspråk i ulike samfunn førte til store samfunnsmessige endringar og framsteg, samtidig som skriftspråka vart ein delar: deltaking i samfunnet føreset bruken av skriftspråket, og bruken av andre språk blir dermed indirekte (eller direkte) undertrykt, med mindre samfunnet aktivt tillét at ein nyttar fleire skriftspråk parallelt. Sjølv om det finst mange døme på slik parallell bruk, er dei unnatak heller enn regel i den store samanhengen.

Utviklinga av boktrykkjarkunsten var fyrste steg på vegen mot massekommunikasjon, med Internett og sosiale media som dei nyaste stega på den vegen. Dei kommunikasjonsbaserte samfunna vi ser i dag, er svært skriftspråksbundne, og bruken av språkteknologi er ein grunnleggjande føresetnad for å kunna delta i samfunnet. Ein kan ikkje delta på eigne premiss og eige språk om språket ikkje er skriftfest, eller om det tilhøyrande skriftspråket ikkje er kodifisert i Unicode[2]. Og sjølv om skriftsymbola i språket er koda i Unicode, er det ikkje sikkert at det finst i tilgjenge-

1 Jf. Ethnologue-statistikk – språk utan skriftspråk er rekna som summen av EGIDS-kategoriane 6a-9, dvs. 4889 av i alt 7105 språk (pr. 4.3.2012). Dette rimar bra med andre kjelder, t.d. føreordet til Thousand Languages, som viser til at kring 2/3 av språka i verda er utan skriftspråk.

2 http://www.unicode.org

lege skrifter for datamaskiner, eller at det finst tastaturoppsett som gjer at ein kan skriva det. Alt dette er språkteknologi som t.d. norsk- og svensktalande tek som sjølvsagt, men som ikkje finst for store delar av språka i verda (Moshagen & Trosterud, 2008).

Ei viktig drivkraft bak arbeidet med samisk språkteknologi er altså å gje det samiske samfunnet dei same verktøya og den same teknologien som finst for majoritetsspråka kring dei, slik at samisk i framtida òg har dei naudsynte språkteknologiske føresetnadene for å bli brukte. Utan språkteknologi blir språkbruken privatisert og utestengd frå det meste av samfunnet, og vil til slutt ikkje eksistera, og då er språka daude.

Domenetap og domenevinning

Ei anna side av denne utviklinga og andre endringar i samfunnet er knytt til dei samfunnsområda, domena, som språket kan bli brukt i og på.

Alle språk har i utgangspunktet vore samfunnsberande, dvs. vore det einaste språket ein har trengt for å klara seg i det samfunnet ein lever i. Men i samband med utviklinga av dei moderne statane, og særleg knytt til utviklinga av nasjonalstaten, har bruken av andre enn det statsberande språket vorte marginalisert, og i dei nordiske landa – og i mange andre land – vorte aktivt undertrykt.

Dette har ført til store domenetap for samisk. Sjølv om den offisielle undertrykkingspolitikken no er avslutta, er samisk i praksis framleis utestengt frå svært mange domene, av ulike grunnar. Ei anna viktig drivkraft for arbeidet med samisk språkteknologi er difor å gje verktøy for å ta tilbake og vinna nye domene, og å ta tilbake tapte talarar.

Sosial status og demografi/stigma

Ein konsekvens av samfunnsendringane som er nemnde over, er at statusen til språka har endra seg kraftig, frå å vera heilt sjølvsagd og naturleg til å bli kraftig stigmatisert som ein konsekvens av undertrykkingspolitikken. I dag er stigmaet borte mange stader, men ikkje over alt, og ikkje for alle.

Sjølv om språkteknologi truleg spelar ei avgrensa rolle når det gjeld sosial status og stigma, så vil han vera ein av mange faktorar som dreg i rett retning: det at samisk har dei same verktøya som t.d. norsk, er eitt teikn på at dei er likeverdige.

Tilhøve mellom majoritet og minoritet

Det er likevel langt att til fullt likeverde mellom samisk og majoritetsspråka. All kommunikasjon mellom språkgruppene skjer på premissa til majoritetsspråka, *på* majoritetsspråka. Det er heile tida minoriteten som må tilpassa seg. Dette avgrensar talet på arenaer der samisk kan bli brukt kraftig, og er eit stort hinder for ein meir utbreidd bruk av samisk.

For å få ein betre balanse mellom språkgruppene burde kommunikasjonen gjerast meir på premissa til minoriteten. Det var slik det ofte fungerte før dei moderne statskonstruksjonane vart bygde, då kommunikasjonen mellom sentralmakt og dei ulike folkeslaga typisk gjekk gjennom tolk. Tolking vil alltid gje informasjonstap, så enno betre er det sjølvsagt om sentralmakta lærer seg språka til dei samfunna ein skal administrera. Og når det gjeld tolking, omsetjing og språklæring, kan språkteknologi vera til hjelp.

Institusjonsbruk

Med slike verktøy vil det på sikt bli mogleg for samiske institusjonar å fungera heilt på samisk sjølv om dei kommuniserer mykje med majoritetssamfunnet. I dag fungerer dei aller fleste heilt på norsk, svensk eller finsk. Om ikkje sametinga kan fungera på samisk, kven kan då gjera det? Kva slags signal gjev det til både det samiske samfunnet og majoritetssamfunnet at sjølve symbola på det samiske folket ikkje opererer på samisk?

eSápmi

eSápmi var eit prosjekt starta i 2002 med sikte på å planleggja og førebu ulike satsingar for å gjera samisk tilgjengeleg og brukande på digitale plattformar, nettopp med utgangspunkt i at dei samiske språka må finnast på slike plattformar for at dei skal bli brukte i framtida. Mykje av det som har vorte gjort seinare vart alt skildra i planane som vart utforma den gongen. eSápmi har såleis vore den språkpolitiske grunnsteinen som alt arbeid med språkteknologi hjå det norske Sametinget har bygt på, og framleis er det prosjekt og teknologi som vart planlagt den gongen som enno ikkje er ferdige. Det gjeld til dømes talesyntese for samisk, der eit prosjekt for nordsamisk er i gang og vil vera ferdig i 2014.

Oppsummering av bakgrunnen for arbeidet med samisk språkteknologi

Som teksten over viser, trengst språkteknologi i vid forstand for at samiske språk skal ha ei framtid og haldast i aktiv bruk. Dei må kunna brukast på datamaskiner og andre digitale plattformar. Det vil vera med på å heva statusen og hjelpa til med å ta att tapte domene og vinna nye. Motsett vil mangel på tilgang til slik teknologi forsterka tendensane til språkdaude.

Status i byrjinga av 2014

Elementær databruk

Teiknsett: Innføringa av Unicode var eit enormt framsteg for alle språk i verda. Ved at det er ein altomfattande standard som ligg til grunn for all moderne tekstkoding på alle digitale plattformar, betyr det at det ikkje lenger er eit problem å koda tekst for databruk, uansett språk, inklusive utdøydde språk. Alle samiske språk er representerte i Unicode, og for språk som enno ikkje er det, finst det standardiserte rutinar for korleis ein skal leggja til nye data og oppdatera standarden. Unicode blir brukt over alt, frå minimale mobiltelefonar til store serversystem, og ein treng ikkje lenger gå ut frå at om ein sender ein tekst frå maskin A til maskin B, så er teksten forvrengd eller øydelagd når han kjem fram. Sjølv om det framleis hender, har det gått frå å vera regel til å vera unnatak.

Skrifter: Unicode er ikkje nok, Unicode er berre ein standard for korleis tekst skal kodast inne i datamaskina. For å visa teksten trengst det skrifter, og at dei tilgjengelege skriftene har alle teikn som trengst. Dessverre er det ein hovudregel at skriftene *ikkje* har dei samiske teikna. Det finst ingen standard for skrifter tilsvarande Unicode for teiknkoding. Det finst i dag i praksis to vegar fram for å skapa fleire skrifter med full samisk dekning: å støtta utviklinga av opne skrifter, og å krevja full samisk dekning i offentlege innkjøpsavtaler av datasystem.

Tastatur: dette vart løyst for Linux, MacOSX og Windows kring 2003, og er ikkje lenger ei sak for desse plattformene. Dessverre har ikkje dei løysingane vorte overførte til dei nye mobile systema Android, iOS, og Windows Phone/RT, og dermed er vi tilbake på rute 1. Det er enno ikkje mogleg å skriva dei samiske språka på mobile system, men Unicode-organisasjonen leier eit initiativ – CLDR[3] – for å samla og halda ved like både tasta-

3 http://cldr.unicode.org/index

turdata (frå versjon 22 av CLDR) og andre data (sjå neste avsnitt). Sjølv om ikkje alle OS er dekte av tastaturoppsett skrivne i dette dataformatet (både iOS og Windows Mobile/RT manglar i lista over støtta OS), vil det vera eit viktig steg framover å leggja til støtte for samiske språk der det manglar.

Sortering og andre grunnleggjande data om språk: Dei aller fleste datasystem føreset visse grunnleggjande data om eit språk for å kunna handtera det skikkeleg, og mangel på slike data fører ofte til at eit språk heller blir fjerna frå eit system enn at det blir inkludert med mangelfull støtte. Data det gjeld er slike ting som sorteringsrekkjefølgje, format for dato og tid, namn på land og språk, osb. CLDR (sjå førre avsnitt) inneheld alle slike data, men berre nordsamisk er dekt i rimeleg grad. Det vil vera eit viktig arbeid i åra som kjem å leggja til data for dei andre samiske språka.

Annan databruk

Sjølv om det er mogleg å skilja mellom det ein kunne kalla *vanleg* bruk og meir *avansert* bruk, så er overgangen mellom dei flytande, og av og til føreset såkalla vanleg bruk at ein i utviklingsarbeidet tek vegen om meir avansert bruk. Heile spekteret er difor samla i dette avsnittet. Inndelinga under fylgjer i staden ei inndeling etter kva språk tekstprodusent og mottakar har.

For fyrstespråksbrukarar

Den overordna målsetjinga er å tilby dei same verktøya og tenestene som for majoritetsspråka, og samtidig tilby verkøty og tenester som gjer at samisktalande vel å skriva på samisk heller enn på eit majoritetsspråk. Til dette trengst det i alle fall stavekontrollar[4], grammatikkontrollar, ordbøker[5], terminologi[6], omsetjingsverktøy[7] og talesyntese.

Talesyntese er viktig av fleire grunnar, både i situasjonar der ein sjølv vanskeleg kan lesa, og i skrivesituasjonar. Ein stor del av den vaksne samiske befolkninga har ikkje fått skriveopplæring, og har difor vanskeleg for å uttrykkja seg i skrift sjølv om dei har samisk som morsmål. Talesyntese vil i kombinasjon med ein stavekontroll og ein grammatikkontroll gjera det mogleg for mange av desse å byrja å uttrykkja seg skriftleg, ved

4 divvun.no
5 satni.org og sanit.oahpa.no
6 http://gtsvn.uit.no/termwiki/index.php/Váldosiidu
7 http://divvun.no/doc/tools/autshumato.html

at dei fyrst nyttar korrekturverktøya til å retta så mange feil som mogleg, og deretter nyttar talesyntesen til å sjekka at teksten høyrest rett ut.

For *nordsamisk* finst det meste, og dei manglande verktøya grammatikkontroll og talesyntese er under utvikling. For *lule- og sørsamisk* finst stavekontrollar, ordbøker og noko terminologi, medan resten av verktøya ikkje er påbyrja. Dei manglande verktøya vil kunna byggja på det arbeidet som blir gjort med nordsamisk, men det er opplagt at det må leggjast ned mykje arbeid for å gjera alle verktøy tilgjengelege for desse språka.

For alle andre samiske språk har det vorte starta arbeid på morfologiske analysatorar og elektroniske ordbøker, og i nokon mon terminologi. Men det meste er ugjort. Dei vil likevel dra stor nytte av arbeidet som blir gjort for nord-, lule- og sørsamisk, i form av at dei kan nytta ein ferdig infrastruktur som vil gje dei ferdige rammeverk og malar for å utvikla alle desse verktøya. Dei treng ikkje finna opp hjulet på nytt, det er nok å fylla det med språkleg innhald (men det er sjølvsagt ein stor jobb, og like stor uansett om det finst 50 eller 50 000 talarar), t.d. kan dei morfologiske analysatorane direkte byggjast som stavekontroll for LibreOffice[8].

For andrespråksbrukarar

Målsetjinga for slike brukarar er dels dei same som for fyrstespråksbrukarar, det vil seia tekstproduksjon på samisk, kanskje med verktøy tilpassa dei behova andrespråkstalarar har. I tillegg er det eit mål å få fleire samisktalande.

Andrespråkstalarar har delvis andre behov enn fyrstespråkstalarar, fordi dei typisk gjer andre feil og har ein annan språkleg bakgrunn. Samtidig er grensa mellom fyrste- og andrespråkstalarar mykje meir flytande og breiare for samisk enn for t.d. norsk. Det finst eit breitt spekter av ulik kompetanse i syntaks, bøying, ordforråd og idiom, og det er ikkje mogleg å setja opp klåre kategoriar basert på språkleg kompetanse. I tillegg til dette kjem det konstante presset frå majoritetsspråka som gjer at sjølv morsmålstalarar ofte har mykje språkleg materiale frå majoritetsspråket med i det samiske språket sitt.

Det er opplagt at ordbøker, terminologisamlingar og omsetjingsverktøy er viktige for andrespråksbrukarar, til liks med ulike skrivestøtteprogram. Dei eksisterande stavekontrollane er sjølvsagt til hjelp for desse brukarane òg, men det har vist seg at andrespråksbrukarar og personar

8 http://divvun.no/libreofficeoxt.html (beta)

som lærer seg samisk innimellom, har problem med å bruka verktøya fordi stavekontrollen tillèt ordformer som er marginale men korrekte, og der brukaren gjer skrivefeil som stavekontrollen godtek fordi skrivefeilen tilfeldigvis er identisk med slike marginale former. Løysinga på dette vil vera å laga stavekontrollar tilpassa ulike grupper av andrespråksbrukarar. Dette har enno ikkje vorte gjort, men rammeverket for å laga slike er stort sett på plass.

Tilsvarande vil det sikkert vera behov for å laga ein grammatikkontroll tilpassa andrespråksbrukarar. Eit slikt arbeid ligg lenger fram, og må venta til den fyrste grammatikkontrollen for morsmålstalarar er ferdig.

I tillegg til verktøy for tekstproduksjon er verktøy for språkopplæring viktige for denne brukargruppa, både for dei som kan litt og dei som kan ingen ting. Det finst ei stor gruppe ungdomar og vaksne som ikkje har fått opplæring i det samiske språket av ulike grunnar, og som er svært motiverte for å læra seg det. I byrjinga av 2014 fanst det språkopplæringsprogram[9] for nordsamisk og sørsamisk, og enklare versjonar for skoltesamisk, enaresamisk og kildinsamisk. Det blir òg utvikla språkopplæringsprogram for ikkje-samiske språk med utgangspunkt i den same plattforma.

For kommunikasjon frå minoritet til majoritet

Den overgripande målsetjinga her er å gjera det mogleg for minoriteten å nytta sitt eige språk mest mogleg i staden for at samisktalande alltid skal byta til majoritetsspråket.

Kommunikasjonen er anten skriftleg eller munnleg. Skriftleg kommunikasjon kan ein dela i to: tekst der detaljane og tekstkvaliteten er viktig, og tekst der det viktigaste er å få med seg hovuddraga i innhaldet. Desse to kommunikasjonstypane krev ulike verktøy.

For at munnleg kommunikasjon skal fungera mellom to språk, og slik at begge partar kan nytta sitt eige språk, krevst det tolk. Det finst enno ikkje språkteknologi som kan nyttast i staden for tolkar, men det finst fleire verktøy som kan vera til hjelp for tolkar. Dette vil typisk vera elektroniske ordbøker og termsamlingar. Som nemnt tidlegare, så finst det slike, men omfanget av ordbøkene og termsamlingane er ikkje stort nok for mange tolkeformål, og det meste av innhaldet er skrive for språkparet nordsamisk-norsk. Det er altså mykje arbeid som står att for alle samiske språk.

9 http://oahpa.no

For skriftleg innhaldsformidling vil det ofte vera nok med maskinomsetjing. Sjølv om resultatet som oftast er langt frå perfekt, er det likevel som regel bra nok til å formidla innhaldet i ein tekst i grove drag. Det finst i dag slik maskinomsetjing frå nordsamisk til norsk bokmål[10]. Det finst ingen ting for dei andre samiske språka.

For å produsera presisjonstekst krevst det andre verktøy. Målsetjinga er at teksten skal kunna produserast på samisk, og at ein så effektivt som mogleg kan omsetja teksten til eitt av majoritetsspråka etterpå. Relevante verktøy er maskinomsetjing, omsetjingsminne, ordbøker, termsamlingar og korrekturverktøy på målspråket.

Det er mogleg å nytta *maskinomsetjing* som eit fyrste steg i ein omsetjingsprosess og retta opp teksten etterpå, men kvaliteten er i dag for dårleg til at ein sparar tid på det, og går som nemnt berre frå nordsamisk til norsk. Over tid vil maskinomsetjinga bli betre og gje meir hjelp i omsetjingsprosessen, og truleg vil det koma fleire språkpar, både frå andre samiske språk og til dei andre majoritetsspråka.

Dei viktigaste verktøya vil vera *omsetjingsminne og parallellkorpus* – som langt på veg er to sider av same sak. Med eit omsetjingsminne vil ein få forslag til omsetjing av setningar basert på tidlegare omsetjingar, og for ein god del byråkratisk tekst med mange repetitive element og liknande tekststrukturar frå år til år burde det vera mogleg å spara mykje tid i omsetjingsprosessen utan at dette går ut over kvaliteten. Eit parallellkorpus inneheld i praksis dei same parallelle setningane, men gjer det mogleg å søkja manuelt i staden for å lita på dei automatiske algoritmane i omsetjingsminnet. Sidan svært mykje av tekstproduksjonen til no har vorte gjort på norsk, finst det nesten ikkje parallelltekstar frå nordsamisk til norsk, og dermed heller ikkje noko omsetjingsminne. Derimot finst det eit norsk-nordsamisk parallellkorpus som det er mogleg å søkja i[11]. Det finst enno ikkje tilsvarande for dei andre samiske språka, men eit nytt prosjekt vil byggja opp parallellkorpus for norsk-lulesamisk og norsk-sørsamisk.

Andre viktige verktøy for ein omsetjar vil vera ordbøker og termsamlingar. Det finst slike for dei fleste samiske språka (jf. fotnote 5 og 6), men omfanget av dei er svært avgrensa for alle andre språk enn nordsamisk. Til sist vil omsetjarane sjølvsagt ha nytte av korrekturverktøy på målspråka, og slike verktøy finst for både finsk, svensk og dei to norske skriftspråka.

---

10 http://gtweb.uit.no/mt/index.php?lang=nno

11 http://gtweb.uit.no/korp/?mode=parallel#parallel_corpora=nob&lang=nb

For kommunikasjon frå majoritet til minoritet

Dei språkpolitiske målsetjingane for denne kommunikasjonsretninga bør vera at kommunikasjonen skjer på minoritetsspråket, av same grunn som diskutert tidlegare: det er minoriteten som er pressa språkleg, og som har alt å vinna på å nytta sitt eige språk. Det er altså viktig for å oppnå ein betre balanse mellom språka.

I denne samanhengen vil det i utgangspunktet alltid vera feil å nytta maskinomsetjing. Maskinomsetjing gjev i praksis alltid dårlegare resultat enn manuell omsetjing, og sidan minoriteten alltid kan majoritetsspråket, vil det å nytta maskinomsetjing fort verka mot sin eigen hensikt – det vil få språkbrukarane til å venda seg bort frå den samiske teksten. Ein annan grunn til å unngå maskinomsetjing sjølv om han skulle bli brukbar, er at omsetjinga alltid vil ha ganske mange drag av originalspråket i seg. Så sjølv om t.d. ein nordsamisk maskinomsett tekst vil vera korrekt nordsamisk, vil det framleis vera ein veldig norskfarga nordsamisk tekst. Dette gjeld òg sjølv om ein språkvaskar teksten etterpå.

Dette betyr at ein bør satsa på manuell omsetjing, med støtte av verktøy som omsetjingsminne, ordbøker, terminologisamlingar og parallellkorpus. Som det vart nemnt i førre avsnitt så finst det eit parallellkorpus frå norsk bokmål til nordsamisk. Det inneheld ca. 156 000 setningspar, frå i all hovudsak administrative tekstar. Tilsvarande finst det eit omsetjingsminne basert på dette korpuset, til hjelp ved omsetjing av slike tekstar. Og det finst ordbøker og termsamlingar, som òg har vorte nemnt før. Men som det alt har vorte klårt, så finst det meste berre for nordsamisk og språkparet norsk bokmål-nordsamisk. Det finst lite eller ingen ting for dei andre samiske språka, og dei potensielle parallellkorpusa er òg så små at dei ikkje vil vera til stor hjelp.

For mange av dei samiske språka vil det difor kunna vera ei brukbar løysing å basera seg på omsetjingane til nordsamisk.

Kommunikasjon mellom minoritetsspråka

Språkpolitisk vil det vera mykje betre om dei samiske språkbrukarane kommuniserer på samisk seg i mellom enn at dei tyr til eitt av majoritetsspråka når dei skal kommunisera på tvers av dei samiske språkgrensene.

For munnleg kommunikasjon trengst det opplæring i nabospråksforståing og ordbøker mellom dei samiske språka. Ingen av delane finst

enno, men eksisterande termsamlingar inneheld noko terminologi for fleire samiske språk, og kan difor vera til hjelp.

For skriftleg kommunikasjon vil settet av verktøy langt på veg vera det same som for kommunikasjon frå minoritetsspråk til majoritetsspråk (dvs. maskinomsetjing, omsetjingsminne, terminologisamlingar og parallellkorpus). Men det er ein viktig skilnad: sidan språka er så like, er det truleg mykje som kan overførast direkte frå eitt språk til eit anna heller enn å omsetjast. Og det er ikkje umogleg å laga maskinomsetjing som vil gjera det raskare å vaska teksten etterpå enn å omsetja han manuelt. Eit fyrste eksperiment med maskinomsetjing frå nordsamisk til sørsamisk har vorte gjennomført, med lovande resultat (sjå Antonsen, Tyers og Trosterud (i kjømda)).

Problema med å nytta maskinomsetjing som vart skildra over, for kommunikasjon frå majoritetsspråk til minoritetsspråk, er heller ikkje til stades i same grad nettopp fordi språka er så like: ein ikkje-neglisjerbar del av leksikonet vil vera felles, mykje av syntaksen er lik, og morfologi og ordstruktur er stort sett lik. Dessutan er det rimeleg å tru at det vil vera betre med eit preg av eit anna samisk språk enn at majoritetsspråket skin gjennom på ulike vis. Dette betyr sjølvsagt ikkje at det enkelt å få eit fungerande system, men resultata frå eksperimentet med nordsamisk-sørsamisk er så lovande at det no startar eit stort prosjekt ved UiT Norges arktiske universitet som skal gå over fleire år og omfattar både studentar, ein doktorgrads-student og ei postdok.-stilling.

Oppsummeringsvis finst det altså enno svært lite språkteknologi og ressursar for kommunikasjon mellom dei samiske språka, men det er arbeid på gang, og det vil truleg vera synlege resultat om nokre år.

For forskarar

Bakom dei fleste av dei omtalde ressursane og verktøya ligg det ulike verktøy og ressursar meir retta mot forskarar, og utvikla av språkforskarar. Det gjeld til dømes slike ting som *morfologisk analyse*. Eit analyseprogram er mest nyttig for forskarar, men er samtidig grunnmuren i all annan språkprosessering, inklusive sluttbrukarverktøya.

Både syntaktisk analyse og analyserte korpus er svært viktige verktøy for både forskarar og til dømes terminologar. I praksis betyr dei tilgjengelege korpusressursane for nordsamisk nærmast eit paradigmeskifte for

forsking på samisk syntaks ved at ein stor del av den samiske tekstproduksjonen er tilgjengeleg på nettet, og slik at ein kan søkja både på ordformer, grunnformer og ulike syntaktiske kontekstar. Etter at prosjektet med talesyntese er avslutta, vil det òg gje verktøy som opnar nye høve til forsking, t.d. kring intonasjon og uttale.

Oppsummering av ressursar og verktøy som finst

Nedanfor er statusen for dei ulike språka oppsummert i tabellform.

	Nordsamisk	Lulesamisk	Sørsamisk	Enaresamisk	Skoltesamisk	Kildinsamisk	Pitesamisk	Umesamis
I Unicode	3	3	3	3	3	3	3	
Har alle bokstavane i minst ein systemfont	3	3	3	3	3	3	3	
Har tastatur for datamaskin	3	2	2	2	2	2	1	
Kan skrivast på mobilsystem	1	1	1	1			1	
Ordliste på mobilsystem								
morfologisk analysator	3	3	3	1	1	1	1	
retteprogram	3	3	3	1	1	1	1	
automatisk orddeling	3	3	3					
grammatikkontroll	1							
elektroniske ordbøker mellom samisk og majoritetsspråk	3	2	3		2	2		
elektroniske ordbøker mellom dei samiske språka	1	1						
terminologisamlingar	3	2	2	2	2			
maskinomsetjing frå samisk til majoritetsspråk	3							
maskinomsetjing frå samisk til samisk	1	1	1					
omsetjingsminne	3							
korpus	3	2	2			1		
parallellkorpus	3	1	1			1		
syntaktisk analyse	3	2	2					
talesyntese	2							
språklæringsprogram	3		3	2	2	2		

Tabell over verktøy og ressursar for dei ulike samiske språka.

Finst ikkje	
Påbyrja / finst for nokre system	1
Under utvikling / finst for dei fleste systema	2
Ferdig versjon / finst for alle system	3

Nykel til fargekodane i tabellen.

Infrastruktur for å byggja ressursar og verktøy

I tillegg til dei verktøya og ressursane som er skildra over, har det vorte bygt opp ein infrastruktur som gjer det mogleg å produsera mange av ressursane og verktøya over, ut frå rådata eller frå eit sett med grunnressursar. Til dømes blir morfologisk analysator (og generator), stavekontrollar og orddelingsprogram i all hovudsak bygde frå den same kjeldekoden, og morfologisk analyse er ein komponent i fleire av dei andre verktøya.

Med ein felles infrastruktur vert det mogleg å produsera langt fleire verktøy for mange fleire språk, sjølv om det grunnleggjande språkarbeidet alltid må gjerast for kvart enkelt språk. I mars 2014 var det 42 språk som det blir jobba aktivt med i denne infrastrukturen, og alle desse språka vil automatisk få fleire av dei nemnde verktøya automatisk når ein utviklar den morfologiske analysatoren.

Slik automatisk gjenbruk er viktig for språksamfunn med svært avgrensa middel og menneskelege ressursar ein kan leggja ned på dette arbeidet. I tillegg blir alle språkteknologiprosjekt oppmoda til å nytta ein så open lisens som mogleg, slik at ein ikkje stengjer ute andre som vil nytta kjeldekoden til andre formål. På den måten vil ein få mest mogleg att for det arbeidet ein legg ned på å utvikla dei språkteknologiske ressursane.

Oppsummering og framtidstankar

Situasjonen for nordsamisk er på mange måtar bra, mange av verktøya og ressursane som har vorte nemnde for at samisk skal kunna fungera heilt samfunnsberande igjen, er på plass, om enn i mindre målestokk enn for andre språk, og mindre omfattande enn det ein treng. Ein kan seia at for nordsamisk er veldig mange av verktøya og ressursane i versjon 1.0, og det trengst mykje arbeid enno for å utvikla dei vidare. På den andre sida har arbeidet med nordsamisk lagt grunnen for tilsvarande arbeid for andre språk, og gjort det mykje lettare for dei å koma i gang med tilsvarande

arbeid. Det same kan seiast om arbeidet med å byggja ein infrastruktur som støttar opp om gjenbruk, felles bruk og eit felles sett med verktøy.

Situasjonen for lule- og sørsamisk er noko dårlegare enn for nordsamisk. Dei viktigaste ressursane og verktøya er på plass, men mykje står att. Dette gjeld særleg korpusressursar, der hovudproblemet er at det finst så få tekstar å leggja inn i eit korpus. På sikt kan ein vona at tilgangen til dei nye verktøya vil stimulera til meir tekstproduksjon som vil gje meir tekst til forskings- og utviklingsarbeid, som igjen vil gje betre verktøy. På den måten kunne ein få ein god sirkel.

For dei andre samiske språka er mykje av arbeidet enno i startgropa. Samtidig blir det arbeidd med dei aller fleste av dei, i større eller mindre grad. Og gjennom den felles infrastrukturen vil alt grunnarbeid som blir gjort for eitt språk, automatisk koma andre språk til gode. Slik sett ser det positivt ut.

Dei to største utfordringane når det gjeld språkteknologi for samiske språk, er å gjera dei praktisk brukbare på mobile system, og mangelen på tekstar. Dei mobile systema er mykje meir lukka enn dei etablerte systema for borddatamaskiner, og har ikkje overført grunnleggjande ressursar for skriving frå bordsystem til mobilsystem. Mobilsystema har difor i praksis vorte ei ny digital sperre for minoritetsspråk, nett som den førre digitale sperra var på veg til å bli rive ned. Sjølv om det no er mogleg å skriva t.d. nordsamisk på iOS-einingar, finst det ikkje støtte for ordlister og ordfullføring. Ein må nytta det finske tastaturet som betyr at ein for alle samiske teikn, må trykkja og halda på ein bokstav for å få fram det samiske teiknet. Berre dei mest ihuga språkbrukarane gjer det.

Mangelen på tekst på samiske språk er ikkje noko som enkelt kan løysast. Men synet på samisk har dramatisk endra seg til det betre over dei siste tjue-tretti åra, og i kombinasjon med dei skrivestøtteverktøya som er utvikla eller er på veg til å bli utvikla, kan situasjonen på sikt bli mykje betre.

Den teknologiske plattforma som har vorte bygt opp og heile tida blir utvikla, gjev eit godt grunnlag for språkteknologiske verktøy og ressursar for dei samiske språka. Samtidig finst det ei positiv og aukande interesse for dei samiske språka både i samfunnet generelt og i ulike finansieringsorgan. Det er difor grunn til å vera optimistisk når det gjeld utviklinga av samisk språkteknologi og verktøy og ressursar for det samiske samfunnet. Til sjuande og sist er det likevel andre faktorar som er avgjerande: språka må haldast i bruk og overførast frå ein generasjon til den neste. På

det punktet kan språkteknologien berre vera eitt av mange tiltak som aukar statusen til samisk og på det viset får fleire til å nytta språka.

Dei fleste samiske språksamfunna er svært små og sårbare, og har vorte hardt råka av språkpolitikken på 1900-talet. Haldningane og kunnskapen har heldigvis endra seg kraftig, men situasjonen er framleis svært kritisk for mange av dei samiske språka. Språkteknologiske verktøy og ressursar er berre ein del av arbeidet med å føra språka vidare. Men med det grunnlaget som no er bygt opp og den aktiviteten som har starta, så finst det håp om at språkteknologi kan gje ei viktig hjelp til dei samiske språka.

Sjur Nørstebø Moshagen er prosjektleder for Divvun-prosjektet ved Universitetet i Tromsø.

Summary

Language technology in all aspects has been a divider among languages since the very invention of writing. Today access to language technology tools that are taken for granted among the majority language speakers can be the factor that ensures a language will continue to be used in the future, and thus survive as a living language. The article goes briefly through the use of language technology among the Sámi languages, and then describes the present state. At the end there's a short look at the future work and possibilities for the Sámi languages and their use of language technology. Three aspects are important for future use: use of open standards, all Sámi resources should be available as open source, and the use of a shared, co-developed infrastructure to ensure that all languages get access to the same support and tool-set.

Litteratur

Antonsen, Lene, Tyers, Francis og Trosterud, Trond (i kjømda): A North Saami to South Saami machine translation prototype, i *Proceedings of the SALTMIL Workshop at LREC2014*

Moshagen, Sjur og Trosterud, Trond, 2008: Datorstöd för samiska och andra minoritetsspråk, i *Tekniken bakom språket*, red. Domeij, Språkrådet, Sverige.