Da forskere ved UiT skapte grunnlaget for en digital samisk skriftkultur

Mun ráhkistan Giellatekno,	***Jeg elsker Giellatekno,***
Divvun-prográmma,	*Divvun-programmet,*
Divvun-speller-demo ja	*Divvun-speller-demo og*
Giellatekno Apertium nuvttá jorgalanprográmma	*Giellatekno Apertium gratis oversettingsprogram*
Giellatekno lea fiidnámus mii lea	*Giellatekno er det fineste som finnes*
Giellatekno lea oavdu	*Giellatekno er et underverk*

fra Siri Broch Johansens bok *Reivvet kommišuvdnii / Brev til kommisjonen* fra 2020.

Tilbakemeldinger i form av dikt som dette er ikke hverdagskost å få for universitetsforskere. Her vil vi fortelle hvorfor vi, som er ei gruppe forskere ved UiT Norges arktiske universitet, kunne lese om arbeidet vårt på denne måten. De samme forskerne fikk i 2012 den nordsamiske samiske språkprisen Gollegiella for arbeidet.

Det var ikke selvsagt at de samiske språkene skulle bli med inn i den digitale tidsalderen. De hadde dårlige odds: Den nordsamiske rettskrivinga hadde 7 bokstaver som ikke fins i norsk. De samiske språka har komplekse bøyingsmønster, med hundrevis av ulike former av samme grunnord og de ulike ordformene laget med hjelp av komplekse prosesser som medfører også endringer inne i ordene. Selv om vi i dag hadde samlet all tilgjengelig samisk tekst, ville relativt vanlige ordformer mangle, fordi tekstene i stor grad er oversettelser av offentlig informasjon.

For å kunne bli med i den digitale revolusjonen, måtte samiske språk inn på en ny arena, fjernt fra både tradisjonelt samisk samfunnsliv og offentlig norsk språkpolitikk. Datarevolusjonen overlot i praksis den globale språkpolitikken til en liten gruppe internasjonale selskaper som i begynnelsen knapt brydde seg om norsk, og slett ikke om samiske språk. Teknologien deres passet også svært dårlig for samiske språk.

Språkteknologien brukes også i oversettelsesprogrammer. Her er en nordsamisk avissak (øverst til venstre) maskinoversatt til lulesamisk og norsk (venstre kolonne) og sørsamisk og enaresamisk (høyre kolonne).

Sett i lys av dette utgangspunktet er situasjonen i dag for de samiske språkene i digital sammenheng overraskende god. Nesten alle verdens språk har få talere, en kompleks

språkstruktur og en svak skriftkultur. Dette gjelder også samisk. Men til forskjell fra *alle andre* språk i samme situasjon, har de samiske språka retteprogram, program for grammatikkontroll, maskinoversetting, syntetisk tale, e-ordbøker, e-læringsprogram og automatisk syntaktisk analyse.

Bak disse brukerprogrammene ligger regelbaserte maskinlesbare modeller av samisk grammatikk. Det teoretiske og matematiske grunnlaget for disse modellene ble lagt på 70- og 80-tallet. Grammatikere viste hvordan regler som simulerte grammatiske endringer kunne bli gjort om til modeller som kunne både analysere og frembringe ordformer. Et viktig senter for dette arbeidet var Helsingfors. Denne innsikten tok vi så med oss til UiT. Der var det politisk vilje til å satse på språkteknologi for urfolksspråk. Det var også store behov: Sametinget hadde en tospråklig administrasjon, og de samiske kommunene skulle ha sakspapir på to språk. I skolene skulle elevene ha lærebøker. Den nordsamiske avisa ønsket å bli dagsavis. Ingenting av dette var mulig uten språkteknologi.

I år er det 20 år siden arbeidet med samisk språkteknologi startet opp ved UiT. Det samiske språkteknologiske miljøet er et samarbeid mellom akademiske lingvister med kunnskap om bygging av språkmodeller, morsmålstalere med filologisk utdanning og programmerere, sammenlagt 10 stillinger. Morsmålstalerne er sentrale i arbeidet, fordi de kan fylle ut det som ikke står i grammatikker og ordbøker, og for samiske språk utgjør det ganske mye.

I arbeidet har vi kombinert kunnskapen vår om samisk språk med arbeidet som hadde blitt gjort i Helsingfors. Samisk grammatikk er på mange felt langt mer komplisert enn finsk, så vi måtte utvide den grammatiske formalismen til å dekke flere fenomen. Vi videreutviklet også teknologien og gjorde den tilgjengelig i flere dataprogram, også åpne program som LibreOffice. Vi har tatt i bruk språkmodellene våre på domener slike modeller ikke hadde blitt brukt før: e-ordbøker, interaktiv e-læring og regelbasert maskinoversetting. F.eks. samarbeider vi nå med den samiske dagsavisa Ávvir for at de også skal kunne tilby leserne avisa på lulesamisk via maskinoversetting.

I dag er det en annen type språkteknologi enn vår som dominerer. Istedenfor å skrive språkmodeller basert på grammatisk kunnskap, er språkteknologi i dag maskinlæring basert på tekstsamlinger i en størrelsesorden som ikke finnes for minoritetsspråk. Den nyeste språkmodellen for norsk baserer seg for eksempel på en tekstsamling som er 12.000 ganger så stort som det som er tilgjengelig for sørsamisk. Et sentralt premiss for maskinlæring er at materialet den baserer seg på, reflekterer språkbruken slik vi vil den skal være. For minoritetsspråk er dette langt fra opplagt. Skriftspråket står svakt, det er stor variasjon både i rettskriving, grammatikk og ordvalg. De tre samiske språkene i Norge fikk nye rettskrivinger så seint som på 1970 og -80 tallet, og det er fremdeles mangler i normeringen. Teknologien vår er dermed fremdeles den eneste mulige for samiske språk.

Et språk får ikke en sterk skriftkultur av ordretteprogram alene, språkteknologi trenges i stadig flere brukerprogram. I tillegg til å lage språkmodeller og ulike program for samiske språk, har vi også laga en generell infrastruktur både for bygging av grammatiske modeller, og for å legge dem inn i ulike program. Denne infrastrukturen tilbyr vi til andre, og den er nå i bruk i større eller mindre grad for 50 språk. I tillegg til 6 samiske språk gjelder det bl.a. kvensk, grønlandsk, plains cree, færøysk, kornisk, meänkieli, haida, ingrisk, komi, komipermjakisk, erzja, moksja, øst- og vestmarisk, livvisk, livisk, tuvinsk, udmurtisk, võro, mansisk og nenetsisk (språkmodellene er åpent tilgjengelige). Felles for alle disse språkene er at de har en kompleks grammatikk, få eller i noen tilfelle ingen tekstressurser, få talere og begrenset tilgang til ressurser for språkteknologisk utviklingsarbeid. Med å utnytte forskings- og utviklingsarbeidet ved UiT, får disse språkene nå tilgang til språkteknologiske løsninger uten å ha like store ressurser til rådighet som vi har hatt for samisk.

Språkteknologi er et nødvendig, men ikke tilstrekkelig vilkår for en skriftkultur. Kunnskapen og viljen blant språkbrukerne til å bruke språket sitt i skrift må også være til stede. Det er ikke alltid tilfelle, og mange av språkene vi har vært med å arbeide med, har en vei å gå før de er i stand til å utnytte mulighetene teknologien gir dem. For de samiske språkene har dette grunnlaget vært til stede. Arbeidet som filologer, lingvister og programmerere ved UiT har gjort for samisk språkteknologi har blitt tatt imot med åpne armer av alle som de siste tiåra har tatt samisk i bruk også som skriftspråk. I løpet av de første sju årene ble de samiske ordretteprogrammene lastet ned over 20.000 ganger. E-ordbøkene for de ulike samiske språka har over 3 millioner oppslag i året. I løpet av en måned oversatte det samisk-norske oversettelsesprogrammat jorgal.uit.no over 28.000 tekster og nesten 2000 nettsider. For et språksamfunn med under 25.000 talere er dette imponerende tall.

Slik gikk det altså til da abstrakte metoder for å modellere grammatisk struktur i maskinlesbar form, kunne gjøre det mulig å etablere en skriftkultur, og også få tilgang til samme type brukerprogrammer som man har tilgang til for norsk og engelsk.

Hilsen medlemmer av Forskerforbundet:
Trond Trosterud, professor i samisk språkteknologi trond.trosterud@uit.no
Lene Antonsen, førsteamanuensis i samisk språkvitenskap lene.antonsen@uit.no
Linda Wiechetek, senioringeniør linda.wiechetek@uit.no
Chiara Argese, overingeniør chiara.argese@uit.no