

Halvårsrapport

1. halvår 2005

10\$Id: half-year-1.xml 4526 2005-09-26 11:40:59Z sjur \$Endeleg versjon

by Sjur N. Moshagen

Eit samandrag av framdrifta i prosjektet det fyrste halve året av prosjektperioden.

Table of contents

1 Planlagt framdrift.....	2
2 Faktisk framdrift.....	2
2.1 Lingvistisk arbeid.....	2
2.2 Prosjektinfrastruktur.....	2
2.3 Korpusarbeid.....	3
2.4 Korrekturprogram.....	3
2.5 Prosjektleiing.....	4
2.6 Oppsummering.....	4
3 Stadnamn.....	5
4 Framlegg til endringar i prosjektplanen.....	5
4.1 Konsekvensar av endringane.....	6
4.2 Konklusjon.....	6

1 Planlagt framdrift

I timeplanen som vart godkjend av styringsgruppa i desember 2004, var desse måla lista opp for fyrste halvår:

- Frist for offentlege tilbod på dei utlyste delane av prosjektet
- Infrastrukturen er ferdig
- Prosjektorganisasjon og –kommunikasjon fungerer tilfredsstillande

2 Faktisk framdrift

Det fyrste punktet av den planlagde framdrifta er enno ikkje gjennomført, dei andre er fullført. I tillegg har vi gjort ein god del som ikkje var ført eksplisitt i den overordna planen, i hovudsak opplæring og lingvistisk arbeid.

Rapporten går gjennom framdrifta for kvar hovudbolc i prosjektarbeidet: lingvistisk arbeid, infrastrukturbygging, korpusinnsamling, korrekturprogram og prosjektleiing, samt ei ålmann oversikt som avslutning.

2.1 Lingvistisk arbeid

Det fyrste halvåret har vi arbeidd med å oppdatera den nordsamiske modellen, byggja ut leksikonet og korrigerer feil. Mot slutten av halvåret har vi òg byrja å sjå på lulesamisk.

Den nordsamiske språkmodellen har vorte mykje betre under dette halvåret, men nokre hol i den lingvistiske modellen står enno att. Når det gjeld den leksikalske dekninga (dvs kor mange ord som blir kjent att i ein vanleg tekst) er det to hovudpunkt som står att: å byggja ut namneleksikonet, og byggja opp korpuset skikkeleg, slik at vi kan testa mot korpuset og leggja inn ukjende ord. Vi har òg identifisert ein del uklære punkt i normeringa av nordsamisk, og har sendt ei liste med spørsmål som bør klårgjerast til Samisk språknemnd.

Vi har òg byrja arbeidet med lulesamisk, med ein gjennomgang av den språklege modellen. Neste steg blir å byggja ut leksikonet til eit reelt leksikon, pr. i dag er det berre eit testleksikon på kring 1000 ord. Vi har byrja å diskutera med Anders Kintel, forfattar av ei ny lulesamisk-norsk ordbok, om å få bruka materialet hans. Aller helst vil vi ha eit kontinuerleg samarbeid med han.

2.2 Prosjektinfrastruktur

Infrastrukturarbeidet har konsentrert seg om desse punkta:

- versjonskontroll
- feildatabase
- internkommunikasjon
- dokumentasjon
- eksternkommunikasjon

Alle disse punkta er no gjennomførte og fungerer tilfredsstillande. Vi har ei heimeside for prosjektet på <http://www.divvun.no/>, der all prosjektdokumentasjon, møtereferat, rapportar m.m. finst tilgjengeleg. Feildatabasen vår finst på <http://giellatekno.uit.no/bugzilla/>, og kven som helst kan bla i databasen, og fylgja med arbeidet med dei registrerte feila (men ein må vera registrert brukar for å kunna leggja inn feil og merknader).

2.3 Korpusarbeid

Arbeidet med å byggja opp eit samisk korpus starta etter at infrastrukturarbeidet var stort sett ferdig. Vi har dels arbeidd med ein intern infrastruktur for å integrera korpusa i utviklingsarbeidet, dels med å finna fram til ein passende lisenstekst som vi kan presentera forfattarar og forlag. Korpusinfrastrukturen er langt på veg ferdig, og er brukande som det er.

Arbeidet med ein korpuslisens/tekstlisens starta med å samanlikna to eksisterande lisensar, den som Universitetet i Oslo bruker, og den som er utvikla ved Helsingfors Universitet. Vi har valt å arbeida vidare med Helsingforsmodellen, i og med at den forenkla ei klår oppdeling i ulike roller og ansvarstilhøve, og som derfor gjer det mogleg å endra på dei praktiske arrangementa seinare utan at ein treng å kontakta teksteigarane (forfattarane og forlaga). Dette arbeidet er ikkje ferdig enno, men nærmar seg slutten. Så snart arbeidet med lisensteksten er ferdig, vil vi gå i gang med å samla inn dei konkrete tekstane.

I arbeidet med lisensteksten har vi samarbeidd med Helsingfors Universitet og ein av forfattarane til den finske originalteksten, Prof. Kimmo Koskenniemi. Dei endelege versjonane av lisensteksten vil bli lagt ut på nettet i lag med bakgrunnsinformasjon. Tanken er at lisensavtalane (det er fire av dei i alt) kan brukast om igjen av andre, eller minste fungera som ein mal som andre kan byggja på og tilpassa sine egne formål. Lisensteksten er tilpassa og justert basert på dei røynslene som Helsingfors Universitet med samarbeidspartar har hatt.

Alle viktige tekstprodusentar har allereie vorte kontakta, og er informert om prosjektet. Dei aller fleste har uttrykt seg positivt eller svært positivt til samarbeid med Divvun-prosjektet, det bør derfor ikkje bli noko problem å få samla i hop eit tilstrekkeleg stort nordsamisk korpus. Særleg bør nemnast at både Min Áigi og Áššu er så interesserte at dei har samla for oss både ukorrigerte og korrekturleste manuskript, Áššu sidan slutten av mai, og Min Áigi i to års tid.

2.4 Korrekturprogram

Noko eigentleg arbeid med korrekturprogram var ikkje planlagt for fyrste halvår, men vi valde å prøva å få ferdig ein demonstrasjon av analysatoren vår i kombinasjon med ein stavekontroll basert på han til konferansen FSMNLP-05 i august/september i Helsingfors. Vi klarte det, og har dermed allereie nådd eitt av måla for andre halvår:

å ha ein alfaversjon av stavekontrollen for nordsamisk ferdig. Stavekontrollen vil bli demonstrert på møtet i styringsgruppa.

2.5 Prosjektleding

Det fyrste halvåret av prosjektet har mykje av tida til prosjektleiaren gått med til å gjera ferdig eit tidlegare påbyrja arbeid for Sametinget: nettstaden og termbasen risten.no. Nettstaden er no ferdig, og vart opna for allmenn bruk 22. juni i år.

Risten.no vil vera til hjelp for Divvun òg, i fyrste omgang ved at ein stor mengde terminologi vil vera tilgjengeleg for utviklinga av Divvun. Slik materialet fanst tilgjengeleg tidlegare, på www.samisk-sr.no, kunne det ikkje brukast utan omfattande arbeid. Den tidlegare termbasen var heller ikkje fleispråkleg (berre tospråkleg nordsamisk-norsk), medan den nye termbasen vil vera ei viktig hjelp for å utvikla lule- og sørsamisk terminologi, og dermed til hjelp for den lulesamiske delen av Divvun, og tilsvarande framtidige, sørsamiske prosjekt.

Det er to punkt i arbeidsplanen som har fått lida pga arbeidet med risten.no: arbeidet med å førebu utlysing av ein offentleg tilbodsrunde for integrering av korrekturverktya, og utarbeidinga av ein detaljarbudsplan. Den daglege drifta av prosjektet har gått som planlagt.

Mangelen av ein detaljarbudsplan har til no ikkje vore eit stort problem, dei fleste oppgåvene har vore klåre og vi har halde oss til den timeplanen som vart lagt til grunn ved prosjektstart. Etter kvart som prosjektet byrjar utvikla testversjonar og involverer både eksterne leverandørar og testarar, vil ei meir detaljert styring vera viktig. Dette gjeld òg samarbeidet med Universitetet i Tromsø, og er noko som allereie er diskutert med det prosjektet. Detaljplanlegginga vil bli starta i nær framtid.

Det er heller ikkje enno noko stort problem at arbeidet med tilbodsteksten har vorte forseinka. Ikkje noko av det arbeidet vi har gjort til no har vore påverka av dette, og heller ikkje arbeidet det nærmaste halvåret treng bli påverka i særleg grad av om vi har ei avtale med eksterne leverandørar eller ikkje. Sjølv om det ville vera bra å ha tilbodsprosessen i gang, kan det òg vera fordelar med å venta: med ei meir moden prosjektgruppe og med røynsla frå å laga den aller fyrste testversjonen er vi betre i stand til å skriva ein god kravspesifikasjon.

2.6 Oppsummering

Divvun-prosjektet hadde ei samling i Kautokeino i mai i år, og gjennomførte ei intern evaluering av prosjektet så langt. Evalueringa inneheldt overraskande lite problem og negative merknader. Vi har fått ei god prosjektgruppe med godt samarbeid, og framdrifta har, med ovannemnde unnatak, vore heilt etter planen. Det fyrste halvåret lovar såleis godt for resten av prosjektperioden.

3 Stadnamn

For å sikra ein god stavekontroll når det gjeld stadnamn, har vi vendt oss til dei nasjonale kartverka i Finland, Sverige og Noreg. Ein god stavekontroll er i denne samanhengen ikkje berre ein som kjenner igjen alle vanlege stadnamn, men òg ein som er i stand til å korrigera frå ei norsk/svensk/finsk namneform til den tilsvarande samiske. For dette formålet treng vi parallellister, lister med stadnamn der både det norske og det samiske (og ev. svenske/finske) er oppført parallelt, og helst alle samiske språk på ein gong. Status for dette arbeidet er pr i dag:

Finland

Vi har fått alle nordsamiske namn og dei parallellførde finske namna utan kostnader.

Sverige

Vi har fått lovnad om å få alle samiske stadnamn. Alle namna er overførde frå gamal til ny ortografi, men ikkje alle namna er klassifiserte som samiske. Når dette klassifiseringsarbeidet er ferdig, vil vi få alle namna. Det er enno ikkje klart om det er mogleg å få namna parallellført med dei tilsvarande svenske/finske namna.

Noreg

Vi har tidlegare fått alle samiske namn, inkl alle lule- og sørsamiske namn, som har vore tilgjengeleg i ny ortografi. Det er ikkje mogleg å få parallellførde namn direkte frå Statens kartverk, men det lar seg gjera som ei betalingsteneste frå Norge Digitalt. Vi har bede Øystein Johannessen i UFD og varamann i styret om å be om ei slik liste med parallellførde namn - UFD, som ein av partane i Norge Digitalt, betaler ikkje noko for å få ut data frå Norge Digitalt.

Konklusjon: det har vore mykje lettare å få stadnamn frå Finland og Sverige enn venta, og inntil vidare heilt utan kostnader for prosjektet. Fram til no er stadnamna det absolutt viktigaste bidraget til prosjektet frå finsk og svensk side.

4 Framlegg til endringar i prosjektplanen

På grunn av arbeidet med termbasen risten.no, føreslår eg nokre endringar i planane for prosjektet, nærmare bestemt for tilbodsspesifiseringa. Tidsfristane er henta frå [NHD](#).

- utvalskriterium og eksakte leveransar blir ferdig til møtet i styringsgruppa i september
- styringsgruppa går gjennom og godkjenner utvalskriteria og leveransane
- ferdig offentleg tilbod blir lagt ut 1-2 månader etter møtet i styringsgruppa (forslag: 20. november)
- frist for innkomne tilbod: minst 52 dagar etter offentleggjering (dersom offentleggjeringa er 20. november, blir fristen tidlegast 11. januar)
- forslag til leverandør blir lagt fram for styringsgruppa på neste møte, som gjer endeleg vedtak.
- kunngjering av tildelt kontrakt: seinast 48 dagar etter frist for innkomne tilbod, dvs. seinast 28. februar om innleveringsfristen for tilboda er 11. januar.

Eit vedtak i januar/februar er kring 4 månader seinare enn opphavleg planlagt, men eg trur det både er bra og riktig at både utvalskriteria og det endelege innkjøpsvedtaket blir forankra i vedtak frå styringsgruppa.

4.1 Konsekvensar av endringane

Tidlegare har eg vurdert at det arbeidet som mest blir påverka av at samarbeidet med leverandøren/leverandørane blir forseinka, er arbeidet med å definera eit nytt leksikonformat. Etter arbeidet med den fyrste stavekontrollen, og ei evaluering av nokre alternativ, både når det gjeld kjerneteknologi og stavekontrollmotorar, har vi ei betre forståing av den variasjonen som finst i feltet og dei krava som ulike teknologiar stiller. Eg trur difor ikkje lenger at ei forseinking av starten på samarbeidet med ein eller fleire eksterne leverandørar får så store konsekvensar. Og sidan all dokumentasjon og alt arbeidet vårt er tilgjengeleg over nettet, er det dessutan lett å invitera alle interesserte partar, inkl. firma, til å kommentera arbeidet vårt og koma med forslag til endringar og betringar føre tilbudsfristen, og uavhengig av kven som får leveransen.

4.2 Konklusjon

Med ein revidert plan som skissert over, burde prosjektet vera i rute i høve til den opphavlege timeplanen om eitt års tid, dvs mot slutten av neste sommar.