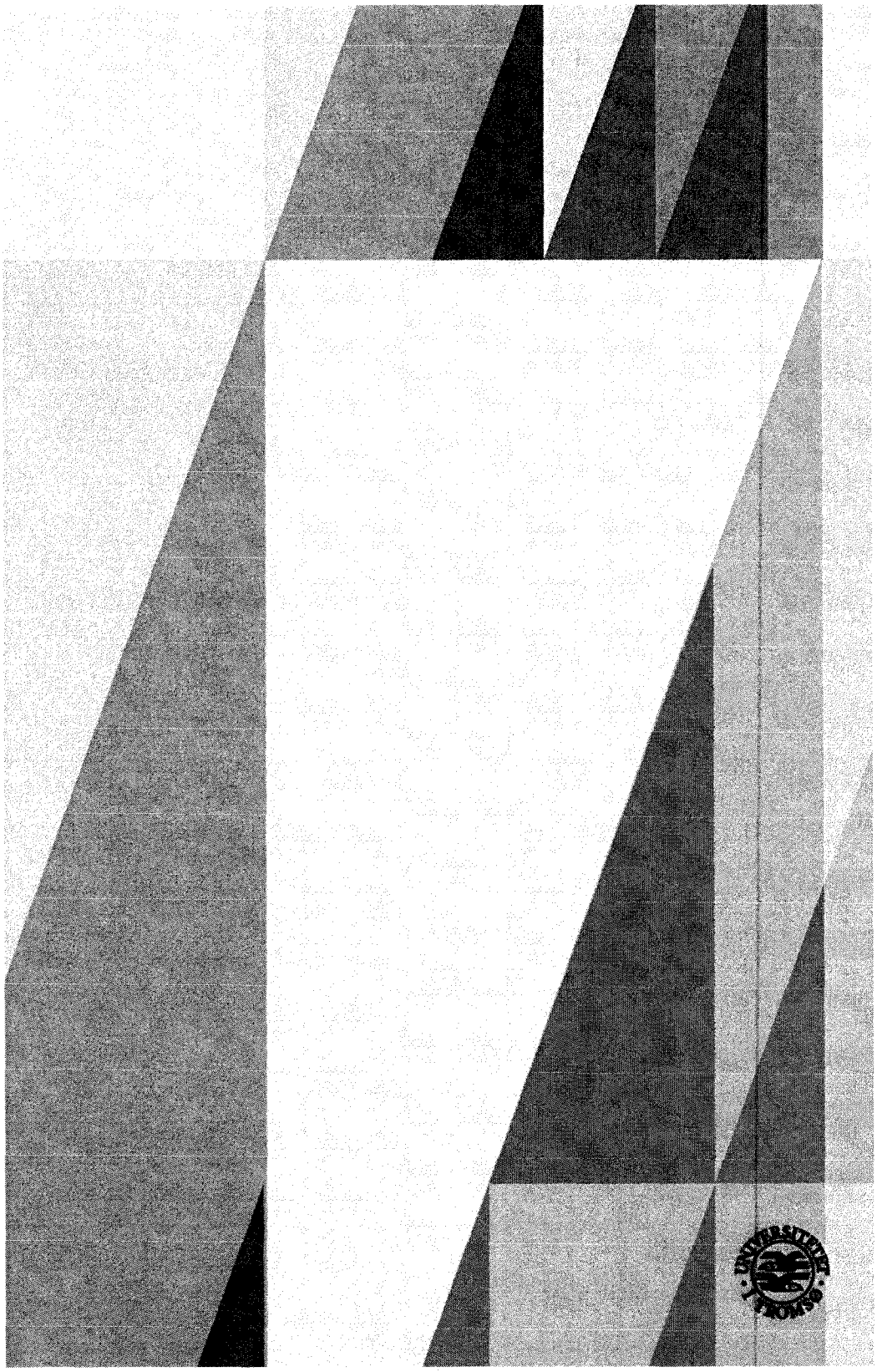
UiT

NORGES ARKTISKE UNIVERSITET

Årsrapport for Divvun 2013

Fakultet for humaniora, samfunnsvitskap og lærarutdanning
Institutt for språkvitskap
Sjur Nørstebø Moshagen / Prosjektleiar / 21.03.2014



Aktivitetar i 2013

Publikumslanseringar

I byrjinga av 2013 leverte vi ei etterlengta oppdatering av den nordsamiske stavekontrollen. Det var fleire år sidan den førre versjonen kom ut, og mange tilbakemeldingar frå brukarane viste tydeleg at det var på tide med ein ny versjon. Han vart lagt ut 6. februar 2013.

I samband med dette arbeidet vart det seinare på våren òg lagt ut ein versjon av alle stavekontrollane som fungerte korrekt for Microsoft Office 2013 under Windows 7 og 8. Microsoft hadde endra visse sentrale innstillingar, og sjølv om retteprogramma i seg sjølv fungerte, så ville dei ikkje fungera under Windows 7 og 8. I samarbeid med utviklarane av den grønlandske stavekontrollen fekk vi alt til å fungera, og den nye installeringspakka er dessutan mykje lettare å distribuera for store organisasjonar.

I samarbeid med Aajege sørsamisk språksenter på Røros har Divvun utvikla ein språklæringsapp for mobiltelefonar i 2013. Appen vart lansert 6. Mars 2014 i ein web-versjon og ein versjon for Android-telefonar. Ein versjon for Apple-telefonar og -nettbrett kjem på våren 2014. Appen rettar seg mot ungdom og unge vaksne som ikkje kan samisk frå før.

Det vart i 2013 gjort mykje arbeid med arvtakaren til risten.no — sátni.org — som er ein webapp for ordbøker og terminologi. Opphavleg hadde vi planar om lansering til sommaren 2013, men det vart ein for stram tidtabell. I staden vart webappen lansert 6. februar 2014, av Sametingspresidenten. Webappen inneheld fleire ordbøker, og all terminologi som finst i Divvun sin termwiki eller som er arbeidd fram av Divvun- og Giellatekno-miljøet (sjå neste avsnitt).

Leksikografi og terminologi

Arbeidet med å henta ut og leggja til rette administrativ terminologi for nordsamisk vart avslutta i 2013. Arbeidet var delt mellom Divvun og Giellatekno, der Divvun stod for korpusinnsamling og tilrettelegging. Materialet finst no tilgjengeleg som ein del av webappen sátni.org.

Det har lenge vore eit stort behov for eit verkty for å redigera og utvikla samisk terminologi. I samarbeid med Helsingfors universitet tok Divvun i bruk ein MediaWiki-basert løysing. Løysinga, som er utvikla av ei gruppe terminologar og datalingvistar ved Helsingfors universitet, vart tilpassa samisk av Divvun, og vi heldt eit innføringskurs i å bruka verktyet hausten 2013. Deltakarane var alle tilsette i Giellagáldu, og dei fleste i det samiske språkteknologimiljøet.

Etter at Sametinget skreiv ein avtale med Anders Kintel om å overta rettane til ordbøkene hans i 2012, starta Divvun og Sametinget i 2013 arbeidet med å bearbeida materialet for betre gjenbruk. Innhaldet i ordbøkene blir omskrive så lite som mogleg, men den leksikografiske strukturen blir gjort eksplisitt og dermed maskinlesbar. Arbeidet vil halda fram i 2014.

Som nemnt over vart nettstaden/webappen sátni.org gjort ferdig i 2013, og lansert 6. februar 2014. Alt leksikografisk og terminologisk materiale for dei samiske språka som Divvun og Giellatekno har tilgang til vil etter kvart bli gjort tilgjengeleg i webappen, i tillegg til andre nettstader og ordboksprogram som språkteknologimiljøa utviklar.

Korpus

Divvun driv eit kontinuerleg arbeid for å samla inn og forbetra korpusdata. I 2013 har arbeidet vore mykje retta mot å forbetra dei data som alt ligg der, ved at konverteringa av materialet har vorte forbetra. I samarbeid med Giellatekno vart det tilgjengelege korpuset lagt ut på nett i eit nytt og forbetra grensesnitt i byrjinga av 2014. Grensesnittet er utvikla av Göteborgs universitet.

Arbeid med ny teknologi

I 2013 vart det starta eller arbeidd vidare med tre ulike prosjekt knytte til ny teknologi. Det fyrste prosjektet er eit langsiktig arbeid med å finna ein god erstattar til den noverande teknologien nytta i stavekontrollane i MS Word. Divvun har i fleire år arbeidd i lag med språkteknologar ved Helsingfors universitet med å byggja ein stavekontroll basert på teknologien deira, og arbeidet nærma seg slutten i 2013. Slike stavekontrollar blir no automatisk bygde for alle språk vi arbeider med, men dei siste delane som trengst for integrering med kontorprogramvare er enno ikkje heilt på plass. Ein alfaversjon for nordsamisk stavekontroll i LibreOffice vart lagt ut 28. februar, og betaversjonar for publikumstesting kjem våren 2014.

Hausten 2013 skaut arbeidet med ein nordsamisk grammatikkontroll fart, og før jul fungerte den fyrste prototypen i LibreOffice. Det er eit stort prosjekt som krev spesialkompetanse, og Divvun har søkt Sametinget om middel til finansiering av delar av prosjektet.

Arbeidet med eit nordsamisk tekst-til-talesystem (kunstig stemme/talesyntese) gjekk inn i ein ny fase i 2013, då Sametinget skreiv under ein avtale med Acapela om å utvikla det ferdige systemet. Divvun bidreg med opptak og språkteknologiressursar. Prosjektet skal vera ferdig mot slutten av 2014.

Infrastrukturarbeid og kvalitetssikring

Arbeidet med å vidareutvikla infrastrukturen vi nyttar for det språkteknologiske arbeidet har gått kraftig framover i 2013, og blir no nytta i dagleg arbeid med over førti språk. Infrastrukturen har òg gjort det lettare å samarbeida med lingvistar og språkteknologar ved andre institusjonar, og i 2013 var eit samarbeid med University of Alberta starta, med fokus på canadiske indianarspråk i nærleiken, fyrst og fremst steppe-Cree. Vi samarbeider frå før med eit prosjekt med sete i Helsingfors som arbeider med uralske språk i Russland.

Med hjelp av infrastrukturen har det vore mogleg å laga stavekontrollar for alle desse over 40 språka, og testa dei ut i LibreOffice, og infrastrukturen gjer det mogleg for lingvistane å fokusera på det språklege arbeidet, utan å måtta bruka store mengder tid på å finna opp hjul og krut på nytt.

Infrastrukturen gjer det òg relativt lett å leggja til testar for ulike delar av dei språklege modellane, og er ein viktig del av arbeidet med å sikra kvaliteten til grammatikkane og ordlistene som blir utvikla for kvart språk.

Andre prosjekt

I 2013 avslutta vi eit prosjekt med å samla inn testdata for skrivefeil på fem ulike nordiske språk (grønlandsk, islandsk, nordsamisk, lulesamisk og sørsamisk). Det innsamla skrivefeilskorpuset har vorte nytta til å testa stavekontrollar for desse språka, og vil vera til hjelp for både språkforskarar og språknormerarar i tida framover. Resultata av prosjektet vil bli presentert på ein workshop hausten 2014. Prosjektet har vorte delfinansiert av Nordplus Sprog.

Divvun har samarbeidd med Giellatekno om maskinomsetjing i 2013. Arbeidet vil bli ført vidare i 2014.

Planar for 2014

Planane for 2014 byggjer langt på veg på aktivitetane i 2013. Det viktigaste blir å vidareutvikla infrastrukturen kring stavekontrollproduksjonen, slik at vi med mindre arbeid kan gje ut oppdateringar oftare. I tillegg vil det allmenne infrastrukturarbeidet vera viktig for å utvikla både verktya vi produserer og kunna gje støtte til alt fleire språk, i lag med standardiseringsarbeid for samisk språkteknologi.

Publikumsretta aktivitet

Til sommaren 2014 vil Divvun koma med oppdateringar for dei lule- og sørsamiske retteverktya. Verktya for Ávvir og forlagssektoren vil bli oppdatert samtidig, og nye verkty for LibreOffice og OpenOffice vil bli tilgjengeleg for publikum for alle språk som finst i infrastrukturen vår.

Arbeidet med eit nordsamisk tekst-til-tale-system (TTS/nordsamisk talesyntese) blir avslutta i 2014. Systemet vil bli eit viktig hjelpemiddel for mange samisktalande, både i skuleverk og i andre samanhengar.

Arbeidet med å trekkja ut terminologi til ei administrativ ordbok for nordsamisk vil bli ført vidare i tilsvarande arbeid med lulesamisk og sørsamisk. Dette arbeidet vil gje delvis andre utfordringar enn for nordsamisk, i og med at det finst så mykje mindre parallelltekst for desse språka.

Arbeid med ny teknologi

I 2014 vil arbeidet med å ta fram ein alternativ teknologi til avløysing for teknologien frå den tidlegare nederlandske underleverandøren bli gjort ferdig, og dei fyrste stavekontrollane baserte på den nye teknologien vil bli lansert, i fyrste omgang for OpenOffice/LibreOffice. Vi vil sjå på røynslene frå denne stavekontrollen før vi vurderer neste steg i arbeidet med avløysarteknologien.

Arbeidet med ein nordsamisk grammatikkontroll vil krevja ein fungerande infrastruktur for grammatikkontrollar, som open kjeldekode og i ulike vertsprogram. Arbeidet som vart starta hausten 2013, og som baserer seg på den finske grammatikkontrollen Voikko, vil bli ført vidare. Målsetjinga er å ha ein fungerande testversjon med avgrensa funksjonalitet i 2014.

Korpus

Divvun vil halda fram med å byggja ut korpusa med så mykje tekst som mogleg.

Infrastrukturarbeid og kvalitetssikring

Arbeidet med å forbetra infrastrukturen – og gjennom det både kvalitet og kvantitet på arbeidet vårt – vil halda fram i 2014, og etter planen skal alle språk vera heilt flytta over til den nye infrastrukturen i år. Infrastrukturen vil fungera som ein plattform for å byggja ut ny funksjonalitet og nye løysingar, og vil gjera det lett å la alle språk få glede av nyutviklingane, som til dømes tekstprosessering for talesyntese, og rammeverk for grammatikkontroll.

Andre prosjekt

Divvun er med i eller har fleire ulike andre prosjekt. Av desse kan nemnast m.a. arbeidet med å få samiske tastatur på mobilsystem (gjerne i lag med retteprogram og ordfullføring). Det vil halda fram i samarbeid med Giellatekno og Sametinget. Dette arbeidet vil bli knytt opp til standardiseringsarbeid for samiske grunnressursar for datasystem i det Unicode.org-leidde

tiltaket CLDR (Common Locale Data Repository), og vil bli gjennomført i samarbeid med det nordiske samiske språksenteret Giellagáldu.

Resultata av det Nordplus Sprog-finaniserte prosjektet med å samla inn skrivefeil og testa stavekontrollar på feilskrivingskorpusa vil bli presenterte på ein workshop på SLTC 2014 i november 2014. Workshoppen blir arrangert m.a. av personar frå Divvun-gruppa.

Eit prosjekt som har vorte nemnt i fleire år, men som det enno ikkje har vore mogleg å gjera noko med, er å bruka maskinomsetjing og omsetjingsminne for å automatisera omsetjing av programvare til samisk. Vi håper å koma i gang med dette arbeidet i 2014, og vil samarbeida både med Giellatekno, Friprog og andre for å gjennomføra eit slikt prosjekt.