Maskinoversetting mellom samiske språk (Forskerprosjekt - SAMISK)

Søknadsnummer: ES518082 Prosjektnummer: 234299

Søker

Prosjektansvarlig

Institusjon / bedrift	UIT-Norges arktiske universitet
Fakultet	Humaniora, samfunnsvitenskap og lærerutdanning
Institutt	Språkvitenskap
Avdeling	
Adresse	Postboks 6050 Langnes
Postnummer	9037
Poststed	Tromsø
Land	Norge
E-post til postmottak	postmottak@hsl.uit.no
Internettadresse	http://www.uit.no
Organisasjonsnummer	970422528
eAdministrasjon	

Administrativt ansvarlig

Fornavn	Jørgen
Etternavn	Fossland
Stilling/tittel	Fakultetsdirektør
Telefon	77644595
E-post	jorgen.fossland@uit.no
Bekreftelse	✔ Søknaden er godkjent av prosjektansvarlig

Prosjektleder

Fornavn	Trond

Maskinoversetting mellom samiske språk (Forskerprosjekt - SAMISK)

Søknadsnummer: ES518082 Prosjektnummer: 234299

Etternavn	Trosterud
Institusjon / bedrift	Universitetet i Tromsø
Fakultet	HSL
Institutt	IS
Avdeling	Giellatekno
Adresse	Breivik
Postnummer	9037
Poststed	Tromsø
Land	Norge
Stilling/tittel	Professor
Akademisk grad	Dr.art.
Ønsket målform	Nynorsk
Telefon	95070140
E-post	trond.trosterud@uit.no

Prosjektinformasjon

Prosjekttittel

Prosjekttittel	Maskinoversetting mellom samiske språk

Prosjektets hovedmål og delmål

Prosjektets hovedmål og delmål	Målet med prosjektet er å lage fungerende program for maskinoversetting fra nordsamisk til andre samiske språk. Ved å oversette maskinelt fra nordsamisk heller enn manuelt fra majoritetsspråket vil hele det samiske språksamfunnet kunne dra nytte av manuelle oversettinger til nordsamisk. Prosjektet vil også gi ny innsikt i samisk komparativ syntaks og ordforråd.

Prosjektsammendrag

Prosjektsammendrag	Målet med prosjektet er å lage fungerende program for maskinoversetting fra nordsamisk til andre samiske språk. Ved å oversette fra nordsamisk til andre samiske språk vil hele det samiske språksamfunnet kunne dra nytte av det arbeidet som blir gjort for nordsamisk. Tekstproduksjon vil bli mulig i en langt større skala enn i dag, f.eks. vil prosessen med å lage skolebøker i alle fag for alle klassetrinn kunne gjøres langt mer effektiv med et maskinoversatt forelegg fra nordsamisk til f.eks. sørsamisk.

Prosjektet vil også gi ny innsikt i samisk komparativ syntaks og ordforråd.

Maskinoversetting er i dag dominert av statistiske modeller (SMT). For samiske språk og andre små språk er dette alternativet ikke mulig. Prosjektet vil derfor bruke lingvistisk basert maskinoversetting (RBMT), bygge videre på arbeid gjort i Tromsø i den siste femårsperioden, og implementere programmene i plattforma Apertium.

Etter en første fase med inventering, innhenting av parallelltekst og komplettering av transferleksikon, vil brorparten av prosjektperioden gå med til komparativ syntaktisk analyse og studier av leksikalsk disambiguering.

Vi ser maskinoversetting mellom nært beslekta minoritetsspråk som en del av det språklige revitaliseringsarbeidet, og dette prosjektet er dermed også relevant for andre minoritetsspråksfamilier.

Plassering

Plassering i Forskningsrådet - tilleggsinformasjon fra søker

Program / aktivitet	SAMISK
Søknadstype	Forskerprosjekt
Delprogram/tema	
Andre relevante programmer/ aktiviteter/prosjekter	Språkteknologi
Disiplin(er)/fagfelt	Språkteknologi
Prosjektnr. v/ tilleggssøknad	
Er relatert(e) søknad(er) sendt Forskningsrådet og/eller annen offentlig finansieringsordning	Nei
Hvis ja, gi nærmere opplysninger	

Framdriftsplan

Prosjektperiode

Fra dato	20140201
Til dato	20170201

Hovedaktiviteter og milepæler i prosjektperioden (år og kvartal)

Milepæler fordelt over prosjektperioden	Fra		Til	
Ansette stipendiater, opplæring	2014	1	2014	2
Infrastruktur, inventering	2014	1	2014	2
Innsamling av tekster	2014	1	2016	2
Oppstart, planlegging	2014	1	2014	1
Teknisk infrastruktur på plass	2014	1	2014	4
Kontrastiv grammatisk analyse	2014	3	2017	1
Evaluering - forbedring - artikkelskriving	2015	2	2017	1
Fungerende betaversjoner i bruk	2016	1	2016	2

Formidlingsplan

Formidlingsplan	Faglig formidling: Prosjektet vil publisere vitenskapelige artikler i relevante faglige fora, og stipendiatene vil skrive sine avhandlinger. Alle lingvistiske ressurser og all kildekode som blir utarbeida i dette prosjektet, vil bli gjort tilgengelig som åpen kildekode, og dermed til nytte også for andre språkforskere og utviklere. Populærvitenskapelig formidling: Kronikker og radioprogram, foredrag på ulike samiske arrangement. Vi vil også tilby nettbasert oversetting via sosiale media. Kommunikasjon med brukere Praktisk formidling til allmenheten: Maskinoversetting som nettsted (http://gtweb.uit.no/mt/), utvidet til url-tjeneste (lim inn url og få nettstedet på et annet språk) Praktisk formidling: Verktøy for profesjonelle oversettere. Formidling til brukere: Profesjonelle oversettere vil få skreddersydde program for datastøtta maskinoversetting basert på maskinoversetting, oversettingsminne og termlister. Det generelle publikum vil kunne oversette tekst via et web-grensesnitt.

Budsjett

Kostnadsplan (i 1000 kr)

	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Personal- og indirekte kostnader	1337	2687	2658	358					7040
Innkjøp av FoU-tjenester									0
Utstyr									0
Andre driftskostnader	25	125	125	25					300
Totalsum	1362	2812	2783	383					7340

Spesifikasjonsfelt	Spesifisert budsjett er vedlagt.

Kostnadssted (i 1000 kr)

	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Næringsliv									0
Instituttsektor									0
UoH-sektor	1362	2812	2783	383					7340
Andre sektorer									0
Utlandet									0
Totalsum	1362	2812	2783	383					7340

Finansieringsplan (i 1000 kr)

	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Egne midler	479	653	560	272					1964
Internasjonale midler									0
Andre offentlige midler									0

	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Andre private midler									0
Søkes Norges forskningsråd	883	2159	2223	111					5376
Totalsum	1362	2812	2783	383					7340

Spesifikasjonsfelt

Person det søkes stipend/stilling for

Fornavn	Etternavn	Fødselsnummer
NN		

Underlag for beregning av stilling

Type stipend	Fra dato (ååååmmdd)	Til dato (ååååmmdd)
Doktorgradsstipend		

	2014	2015	2016	2017	2018	2019	2020	2021
Stillingsprosent								

Underlag for beregning av utenlandsstipend/gjesteforskerstipend

Institusjon / bedrift

Sted

Land

Reiser med familie

Reiseutgifter

Periode

Fra dato (ååååmmdd)

Til dato (ååååmmdd)

Fornavn	Etternavn	Fødselsnummer

NN2

Underlag for beregning av stilling

Type stipend	Fra dato (ååååmmdd)	Til dato (ååååmmdd)
Postdoktorstipend		

	2014	2015	2016	2017	2018	2019	2020	2021
Stillingsprosent								

Underlag for beregning av utenlandsstipend/gjesteforskerstipend

Institusjon / bedrift	Reiser med familie	Reiseutgifter
Sted		
Land		**Periode**
		Fra dato (ååååmmdd)
		Til dato (ååååmmdd)

Fornavn	Etternavn	Fødselsnummer

Underlag for beregning av stilling

Type stipend	Fra dato (ååååmmdd)	Til dato (ååååmmdd)
Ikke valgt		

	2014	2015	2016	2017	2018	2019	2020	2021
Stillingsprosent								

Underlag for beregning av utenlandsstipend/gjesteforskerstipend

Maskinoversetting mellom samiske språk (Forskerprosjekt - SAMISK)

Søknadsnummer: ES518082 Prosjektnummer: 234299

Institusjon / bedrift		Reiser med familie	Reiseutgifter
Sted			
Land			**Periode**
			Fra dato (ååååmmdd)
			Til dato (ååååmmdd)

Søkes Norges forskningsråd (i 1000 kr)

	2014	2015	2016	2017	2018	2019	2020	2021	Sum
Studentstipend		100	100						200
Doktorgradsstipend	858	967	999	86					2910
Postdoktorstipend		967	999						1966
Gjesteforskerstipend									0
Utenlandsstipend									0
Forskerstilling									0
Timebasert lønn inkl. indirekte kostnader									0
Innkjøp av FoU-tjenester									0
Utstyr									0
Andre driftskostnader	25	125	125	25					300
Søkes Norges forskningsråd	883	2159	2223	111					5376

Samarbeidspartnere

Samarbeidspartnere som skal delta i prosjektet med faglige og/eller økonomiske ressurser

Vedlegg

Prosjektbeskrivelse

Filnavn	sme2smX.pdf
Referanse	ES518082_001_1_Prosjektbeskrivelse_20130904

Curriculum vitae (CV) med publikasjonsliste

Filnavn	Trosterud_CV.pdf
Referanse	ES518082_002_1_CV_20130904

Karakterutskrifter (Doktorgrads- og studentstipend)

Filnavn	
Referanse	

Fageksperter

Filnavn	
Referanse	

Anbefaling og invitasjon

Filnavn	
Referanse	

Bekreftelse på samarbeidspartnere

Filnavn	
Referanse	

Annet

Filnavn	Budsjett Trosterud NFR.pdf
Referanse	ES518082_010_1_Annet_20130903

Maskinoversetting mellom samiske språk

1. Bakgrunn og kunnskapsstatus

1.1 Maskinoversetting

Maskinoversetting som fagfelt har de siste 10-15 åra vært dominert av statistisk basert maskinoversetting (SMT), særlig ettersom raskere datamaskiner og mer tilgang til parallell tekst har gjort SMT bedre. Metoden er mest kjent via tjenesten *Google Translate*, men den er også i bruk i andre sammenhenger. Alternativet til SMT er regelbasert maskinoversetting, RBMT, som tidligere var det dominerende paradigmet innafor maskinoversetting. SMT fungerer best for språkpar der språka har lite morfologi (invariante ordformer som opptrer ofte i tekst), og der det er tilgang til store mengder med setningsparallellisert tekst (ideelt over hundre millioner ord, større datagrunnlag gir bedre oversetting). Med parallelltekst fra mange domener vil SMT også være robust og godt egnet til oversetting for å forstå innholdet i fremmedspråklig tekst.

Som vist bl.a. av Koehn 2005 har SMT problemer med språk med rik morfologi, særlig gjelder det språkpar der *målspråket* har rik morfologi. SMT er også avhengig av store parallellkorpora, helst over 10 millioner ord, gode SMT-språkpar har langt mer. For språkpar som f.eks. nordsamisk - sørsamisk er det urealistisk å forvente seg parallellkorpora på så mange hundre tusen ord i overskuelig framtid, og også disse vil være oversatt fra et tredjespråk.

SMT er også mindre egnet til maskinoversetting for tekstproduksjon. Typiske scenarier vil være tekst innenfor et kontrollert domene, der stadig nye tekster over samme tema skal bli oversatt for publikasjon, og der det er viktig å være i stand til å rette feil og stole på et forbedret resultat i en iterativ prosess. For SMT eksisterer ingen slik "feilrettingsprosedyre". Det er ingen garanti for at term T skal bli oversatt til X og ikke til Y i kontekst Z neste gang en tekst oversettes. Man kan bare gi systemet mer parallelltekst og håpe på det beste.

Sett fra et lingvistisk perspektiv er et SMT-system en svart boks. Vi kan åpne den, men det eneste vi vil finne inni vil være endeløse tallrekker. RBMT baserer seg på lingvistiske analyser, og er dermed relevant for en lang rekke språkvitenskapelige spørsmål, noe vi går nærmere inn på i kapittel 2.

1.2 Maskinoversetting ved UiT

Giellatekno, Senter for språkteknologi ved UiT, har siden 2001 utarbeidet grammatiske analyseprogram for samiske og andre sirkumpolare språk. Fra 2004 har vi samarbeidet tett med *Divvun*-gruppa som både har bygd opp grunnressursene sammen med Giellatekno og har laget stavekontroller basert på grunnressursene. Divvun blei en del av UiT i 2011.

Etter hvert som vi har fått grunnleggende analyseprogram på plass har det blitt mulig å gjennomføre mer avanserte forskingsprosjekter. Ett slikt område er maskinoversetting. Fra og med 2008 har vi arbeidet med maskinoversetting, med

følgende resultat:

- Vi har vurdert mulighetene for å bruke SMT og RBMT mellom nord- og lulesamisk (Tyers, Wiechetek & Trosterud 2009), og sett på grunnlaget for leksikalsk disambiguering (Wiechetek, Tyers & Omma 2010)
- Vi har laga en prototype for et maskinoversettingsprogram fra nordsamisk til sørsamisk, der målet er at det skal bli bra nok til at det vil være nyttig som kladd for oversettere. Prototypen inneholder ordforråd for skoleadministrative tekster, og vi er nå i gang med å foreta en evaluering av programmet innafor dette domenet. De første resultatene av arbeidet og evalueringa skal legges fram på en internasjonal konferanse *Oovtâst* i Inari i Finland 25-27. september 2013.
- Vi har deltatt i utviklinga av et maskinoversettingssystem mellom nynorsk og bokmål, dette er sannsynligvis det nest mest brukte MT-systemet i Norge, etter Googles, med oppunder 10000 oversatte tekster i uka (Unhammer & Trosterud 2009).
- Vi har laga et maskinoversettingsprogram fra nordsamisk til norsk, der målet er at norskspråklige skal kunne forstå nordsamisk tekst (Trosterud & Unhammer 2013). Programmet er tilgjengelig online (gtweb.uit.no/mt/) og i daglig bruk.
- Vi har også laga en prototype for oversetting fra finsk til nordsamisk, basert på en grammatisk analysator for finsk og på Giellateknos finsk-samiske ordbok. Programmet danner basis for MT-delen av et felles norsk-estisk forskingsprosjekt (*samest: Sámi - Estonian language technology cooperation Similar languages, same technologies*), som er finansiert over et norsk-estisk forskingsprogram og skal gå over to og et halvt år (2013-15), der forskere i Tromsø og Tartu skal arbeide med maskinoversetting fra finsk til nordsamisk og estisk, og med interaktiv estisk språkopplæring for russiskspråklige, basert på infrastrukturen for det samiske læringsprogrammet *Oahpa*.
- *Linda Wiechetek* som har vært ansatt som doktorgradsstipendiat ved Giellatekno, er i ferd med å avslutte doktorgradsarbeidet sitt (*Not so shallow machine translation for North Sámi*), der temaet er bruk av valensinformasjon for å forbedre maskinoversettinga.
- Giellatekno samarbeider med russiskmiljøet ved UiT om både maskinoversetting og Intelligent Computer-Assisted Langugage Learning (ICALL). UiT har nettopp ansatt *Francis Tyers* i ei postdoktorstilling for en treårsperiode, for å arbeide med norsk-russisk, men også med samisk og finsk. Tyers er sentral i utviklinga av RBMT-plattforma Apertium (wiki.apertium.org), og han har også erfaring både med maskinoversetting for morfologirike språk generelt, og med samiske språk.
- I løpet av Giellateknos samarbeid med Apertium har vi introdusert to-nivåmorfologi og føringsgrammatikk (Apertium inneholdt tidligere bare enkle transdusere for affiksering, og statistiske modeller for disambiguering), og dermed åpnet opp for regelbasert maskinoversetting også av morfologisk komplekse språk. Arbeidet vårt har dermed gitt viktige bidrag til maskinoversetting generelt.

Senter for samisk språkteknologi i Tromsø har nå kommet til et punkt der det er naturlig å satse mer på maskinoversetting. Vi har allerede et godt utgangspunkt, og den kommende tre-fire-årsperioden vil maskinoversetting bli et viktig arbeidsfelt for flere språkmiljø i Tromsø, med norsk-russisk, finsk til nordsamisk og estisk, i tillegg til arbeidet beskrevet i denne søknaden. I tråd med erfaringene våre så langt vil vi

bruke regelbasert maskinoversetting (RBMT), med den åpne plattformen Apertium.

1.3 Regelbasert maskinoversetting

Maskinoversettingsprosessen i de systemene vi er i ferd med å bygge opp skjer på følgende måte:

1. Etter deformattering blir input-tekst analysert morfologisk (med endelige tilstandsautomater, (jf. Beesley & Karttunen 2003) og syntaktisk (med føringsgrammatikk, jf. Bick 1999), og resultatet er morfologisk og syntaktisk tagget lemmatisert tekst.
2. Lemma blir slått opp i et *transferleksikon*, og får målspråksekvivalenter. I tilfelle leksikalsk tvetydighet er det kontekstavhengige regler som velger korrekt oversetting.
3. Et sett av *transferregler* gjør de syntaktiske og morfologiske endringene som skal til for å gå fra kilde- til målspråk. Fra nord- til sørsamisk må for eksempel substantivets Gen.Sg. i tallordsfraser endres til Nom.Pl, og leddstillinga SVO må endres til SOV.
4. Til slutt *genereres* ordformene i målspråksteksten, ved hjelp av målspråkslemma og (evt. revidert) grammatisk analyse.

1.4 Komparativ samisk grammatikk ved UiT

For å være i stand til å skrive leksikalske og grammatiske transferregler mellom ulike samiske språk må lingvisten ha kjennskap til forskjellene språka i mellom. Mange av disse er godt kjent, og behandlet i litteraturen (også i Tromsø, se f.eks. Trosterud 1994, 1996). Andre, mer subtile forskjeller, vil dukke opp i sammenheng med arbeidet med maskinoversettingsprogrammet.

Regelbasert maskinoversetting representerer kvintessensen av tekstbasert språkteknologi, og krever både morfologisk og syntaktisk analyse, leksikalske ressurser, disambiguering av leksikalsk flertydighet, og språkgenerering. Samtidig har det også store samfunnsmessige konsekvenser.

2. Problemstillinger, hypoteser og metodevalg

Prosjektets overordna mål er å utarbeide fungerende program for maskinoversetting mellom samiske språk, med *nordsamisk* som kildespråk og *andre samiske språk* som målspråk. Cirka 80 % av alle samisktalende snakker nordsamisk (Magga 2012: 5), som dermed kan tjene som et *nav* for maskinell oversetting av tekster til andre samiske språk.

De samiske språka er grammatisk komplekse og har relativt fri ordstilling, særlig for konstituenter som ikke er argument av verbet. Vi vil undersøke hvorvidt RBMT samiske språk i mellom vil være i stand til faktisk å gjøre det lettere å produsere tekst på ulike språk. Mer generelt vil et eventuelt positivt resultat vise at maskinoversetting fra slike nav-språk kan være et alternativ til den modellen som dominerer internasjonalt, med bare ett nav (engelsk), og maskinoversetting mellom alle andre språk som relé-oversetting via engelsk.

Vi har som hypotese at det er mulig å lage effektive maskinoversettingssystem for minoritetsspråk med få og begrensede ressurser, av en slik kvalitet at det vil være mulig å oppnå substansiell besparing i oversettingsprosessen. I en situasjon der SMT er bortimot totalt dominerende, er dette i seg selv et resultat relevant for diskusjonen om maskinoversetting.

Ut over hovedmålet knyttet til selve maskinoversettingsprosessen har vi ulike delmål og delhypoteser for prosjektet:

2.1 Hypoteser for den grammatiske forskinga involvert i prosjektet

- Vi vil utforske samisk komparativ syntaks. Vi har en hypotese om at de ulike samiske språka blir påvirka av sine respektive majoritetsspråk i retning bort fra det felles vesturalske utgangspunktet. I tidligere arbeid har vi vist at det er syntaktiske forskjeller i bruken av pre- og postposisjoner mellom nordsamisk skrevet av forfattere fra finsk og norsk side av grensa (Antonsen, Janda og Bals Baal 2012). I hvor stor grad det skjer, og i hvor stor grad nordsamisk (som står i en mellomposisjon mellom finsk og skandinavisk innflytelse) skiller seg fra nabospråka sine, er et spørsmål vi vil prøve å svare på. Vi spår også større syntaktiske skiller fra nordsamisk og vestover enn fra nordsamisk og østover, dette prosjektet vil i større grad enn før være i stand til å karakterisere disse skillene presist. Tradisjonelt er avstanden mellom nord- og lulesamisk sett på som den minste mellom de ulike samiske standardspråka, arbeidet i dette prosjektet kan bidra til å korrigere dette bildet.
- Vi vil utforske terminologiske og leksikalske paralleller og forskjeller mellom de samiske språka. Det er et uttalt mål fra språkrøktshold at det skal etableres felles terminologi, i hvor stor grad det lykkes vet vi lite om, på grunn av en nesten total mangel på leksikalske ressurser mellom samiske språk. Dette prosjektet vil innebære utarbeiding av store intra-samiske ordlister, og dermed gjøre det mulig å undersøke leksikalsk avstand innad i den samiske språkfamilien. Hypotesen vår er at de ulike samiske språka i liten grad skjeler til hverandre i terminologiarbeidet, en systematisk gjennomgang av transferleksika mellom ulike samiske språk vil gi et nøyaktig svar på i hvor stor grad termdanningsprosessen skjer parallelt.

2.2 Hypoteser for de samfunnsmessige konsekvensene av prosjektet

- Vi tror at fungerende maskinoversettingsprogram fra nordsamisk til andre samiske språk vil kunne effektivisere tekstproduksjon på flere samiske språk, og dermed gjøre det mulig å opprettholde en samiskspråklig offentlighet for hele det samiske språksamfunnet. Hypotesen vår er altså at maskinoversetting vil ha merkbare konsekvenser for tekstproduksjon, og at kostnadene for å få f.eks. full dekning av sørsamiske skolebøker for alle klassetrinn vil gå drastisk ned. I forbindelse med det leksikografiske arbeidet ved Divvun og Giellatekno har vi faste rutiner for innsamling av samiskspråklig tekst fra internett, og vi vil dermed på en direkte måte kunne måle utviklinga av samisk tekstproduksjon.
- Vi tror maskinoversettingsprogram og ordbøker mellom samiske språk vil kunne styrke kontakten ulike samiske språksamfunn i mellom. I dag er den samiske samfunnsdebatten fragmentert mellom ulike språkgrupper, bedre tilgang til den nordsamiskspråklige debatten for andre samiskspråklige vil styrke både det samiske språksamfunnet og den pansamiske samfunnsdebatten.

- Vi har som hypotese at maskinoversetting fra nordsamisk til sør- og lulesamisk vil kunne gi opphav til en nordsamiskpåvirket sør- og lulesamisk tekst, heller enn en norskpåvirket.

3. Prosjektplan, prosjektledelse, organisering og samarbeid

Ved oppstart av prosjektet vil vi lyse ut de relevante stillingene. Valg av målspråk vil være avhengig av hvilke stipendiater som vil være best kvalifisert. Ved Giellatekno har vi allerede infrastruktur og påbegynt arbeid for program for oversetting fra nordsamisk til sør- og lulesamisk, så disse språkpara peker seg ut. Men også skoltesamisk er et mulig språkpar for prosjektet, avhengig av utarbeiding av grammatiske grunnressurser vil det også være mulig å inkludere de andre samiske språka.

4. Budsjett

Vi søker om et doktorgradsstipend, et postdoktorstipend på to år, og to studentstipend. Vi vil lyse ut stipendiatstillingene, og valg av målspråk (de mest aktuelle er sørsamisk, lulesamisk og skoltesamisk) vil være avhengig av hvilken søker som er best kvalifisert. Vi vil også lyse ut to studentstipend. Planen er å skape et sammensatt fagmiljø, der forskerne og studentene innafor dette prosjektet i samarbeid med de andre som arbeider med maskinoversetting i Tromsø til sammen kan utgjøre et fruktbart miljø. Vi søker også om en viss grad av programmeringsressurser.

Størstedelen av finansieringen av det nødvendige arbeidet med infrastruktur vil være egenfinansiering fra UiTs side, og inngå som en del av Giellatekno sitt ordinære arbeid. Ut over det vil Francis Tyers, som er postdoktor i russisk språkvitenskap, bruke tid på infrastrukturen i dette prosjektet, i og med at dette prosjektet vil dele infrastruktur med andre maskinoversettingsprosjekt. Trond Trosterud vil bruke ca. halve forskingsinnsatsen sin (25 % stilling) til maskinoversetting i hele perioden. Både Lene Antonsen og Ciprian Gerstenberger vil delta, med tilsammen opptil ett årsverk i løpet av treårsperioden.

5. Strategisk forankring og sentrale perspektiver

5.1 Strategisk forankring

Samisk relevans: Prosjektet vil innebære økt forsking på grammatiske og leksikalske aspekt både ved nordsamisk, men særlig ved andre samiske språk enn nordsamisk. Det vil på en merkbar måte stimulere produksjon av tekst for andre samiske språk enn nordsamisk.

Sirkumpolær relevans: Prosjektet har direkte overføringsverdi til andre sirkumpolare språk, f.eks. maskinoversetting fra grønlandsk til iñupiaq, eller fra komi til komipermjakisk. Ved Giellatekno har vi allerede samarbeid på gang med andre for å lage grunnressurser for slike språk.

Internasjonalt: Prosjektet bidrar til norsk maskinoversetting. Et aktuelt språkpar er russisk - norsk.

5.2 Samfunnsmessig relevans

For de mindre samiske språksamfunnene vil terskelen for å få tilgang til lærebøker og andre samiske tekster på eget språk bli mindre. En sideeffekt av prosjektet vil være intra-samiske ordbøker, dette vil styrke det pansamiske samkvemmet.

5.3 Miljøkonsekvenser og etikk

Vi vil i dette prosjektet følge arbeidsmåten som Giellatekno har hatt i mange år: dokumenter og notater legges inn i et felles versjonkontroll-system i en felles datamaskin og dessuten ut på internett, istedenfor å skrives ut på papir. Vi holder de aller fleste møtene ved hjelp av web-kamera og mikrofon og samskrivingsprogrammer, istedenfor å reise.

I arbeid som kan ha konsekvenser for et minoritetsspråk, er det viktig å ha nært samarbeid med språkbrukerne og deres språknormeringsorganer. Staben i og rundt *Giellatekno* er i stor grad sjøl aktive medlemmer av språksamfunnet, og vi har også samarbeid med Giellagáldu, det nordiske ressurssenteret for de samiske språk, opprettet av sametingene i Norge, Sverige og Finland.

5.4 Rekruttering av kvinner, kjønnsbalanse og kjønnsperspektiv

De aktuelle kandidatene til PhD- og postdoktorstipend er kvinner. Språklig revitalisering er tradisjonelt et område dominert av kvinner, og dette prosjektet vil styrke det fagområdet.

6. Kommunikasjon og formidling

6.1 Formidlingsplan

- Faglig formidling: Prosjektet vil publisere vitenskapelige artikler i relevante faglige fora, og stipendiatene vil skrive sine avhandlinger. Alle lingvistiske ressurser og all kildekode som blir utarbeida i dette prosjektet, vil bli gjort tilgengelig som åpen kildekode, og dermed til nytte også for andre språkforskere og utviklere.
- Populærvitenskapelig formidling: Kronikker og radioprogram, foredrag på ulike samiske arrangement. Vi vil også tilby nettbasert oversetting via sosiale media.

6.2 Kommunikasjon med brukere

- Praktisk formidling til allmenheten: Maskinoversetting som nettsted (http://gtweb.uit.no/mt/), utvidet til url-tjeneste (lim inn url og få nettstedet på et annet språk)
- Praktisk formidling: Verktøy for profesjonelle oversettere.
- Formidling til brukere: Profesjonelle oversettere vil få skreddersydde program for datastøtta maskinoversetting basert på maskinoversetting, oversettingsminne og termlister. Det generelle publikum vil kunne oversette tekst via et web-grensesnitt.

7. Utlysningsspesifikke tilleggsopplysninger

Relevant forskning gjort av miljøet rundt Giellatekno

- Antonsen, Lene, Laura Janda & Biret Ánne Bals Baal 2012: Njealji davvisámi adposišuvnna geavahus. [English summary: The use of four North Saami adpositions.] – *Sámi dieđalaš áigečála* 2012 (2) s. 7–38.
- Antonsen, Lene & Trond Trosterud 2010: Manne dihtor galgá máhttit grammatihka? (Why the computer should know its Sami grammar.) *Sámi Dieđalaš Áigečála* 2010 (1) s. 3–28. Romsa – Guovdageaidnu.
- Antonsen, Lene, Linda Wiechetek & Trond Trosterud 2010: Reusing Grammatical Resources for New Languages. In *Proceedings of the International conference on Language Resources and Evaluation LREC* 2010. p. 2782–2789. ISBN 2-9517408-6-7. Stroudsburg: The Association for Computational Linguistics.
- Antonsen, Lene & Trond Trosterud 2011: Next to nothing – a cheap South Saami disambiguator. *NEALT Proceedings Series* 2011. Volum 14 [10].
- Brandt, M. D, H. Loftsson, H. Sigurþórsson & Francis M. Tyers 2011: "Apertium-IceNLP: A rule-based Icelandic to English machine translation system". Proceedings of the 15th Annual Conference of the European Association for Machine Translation, EAMT11, pp. 217–224.
- Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. & Tyers, F. M. 2011: "Apertium: a free/open-source platform for rule-based machine translation. Machine Translation 24(1) pp. 1–18.
- Otte, P. & Francis M. Tyers 2011: "Rapid rule-based machine translation between Dutch and Afrikaans". Proceedings of the 15th Annual Conference of the European Association for Machine Translation, EAMT11, pp. 153–160.
- Susanto, R. H., Lasarati, S. D. & Francis M. Tyers 2012: "Rule-based machine translation between Indonesian and Malaysian". Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing at the International Conference on Computational Linguistics, COLING2012 pp. 191–200.
- Toral, A., Ginestí-Rosell, M. & Francis M. Tyers 2011: "An Italian to Catalan RBMT system reusing data from existing language pairs". Proceedings of the Second International Workshop on Free/Open- Source Rule-Based Machine Translation. pp. 77-81
- Trosterud, Trond 1994: Auxiliaries, Negative Verbs and Word order in the Sami and Finnic Languages. Ago Künnap (ed.): *Minor Uralic Languages: Structure and Development*. Tartu 1994. pp.173-181.
- Trosterud, Trond 1996: Die südsamische Wortfolge als eine Kombination der deutschen und marischen Wortfolgen analysiert. Lars Gunnar Larsson (Hrsg.): Lapponica et Uralica. 100 Jahre finnisch-ugrischer Unterricht an der Universität Uppsala. *Studia Uralica Upsaliensia* 26: 103-112. Uppsala
- Trosterud, Trond & Kevin Brubeck Unhammer 2013: Evaluating North Sámi to Norwegian assimilation RBMT. In: Cristina España-Bonet and Aarne Ranta (eds.) *Free/Open-Source Rule-Based Machine Translation* pp. 13–26.
- Trosterud, Trond & Linda Wiechetek 2007: Disambiguering av homonymi i nord- og lulesamisk. Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007. Doaimmahan Jussi Ylikoski ja Ante Aikio. *Mémoires de la Société Finno-Ougrienne* 253, pp. 401-422.
- Tyers, Francis M., Washington, J. N., Salimzyanov, I. & Batalov, R. 2012: "A prototype machine translation system for Tatar and Bashkir based on

free/open-source components". Proceedings of the Turkic Languages Workshop at the Language Resources and Evaluation Conference, LREC2012, pp. 11–14.
- Tyers, Francis M., Linda Wiechetek & Trond Trosterud 2009: Developing Prototypes for Machine Translation between Two Sami Languages. *Proceedings of the 13th Annual Conference of the EAMT*, pp. 120-127.
- Unhammer, Kevin & Trond Trosterud 2009: Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. J.A. Pérez-Ortiz, F. Sánchez-Martínez, F.M. Tyers (eds.) *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, p. 35–42 Alacant, Spain, November 2009. http://hdl.handle.net/10045/12025.
- Washington, J. N., Ipasov, M. & Francis M. Tyers 2012: "A finite-state morphological analyser for Kyrgyz". Proceedings of the 8th Conference on Language Resources and Evaluation, LREC2012, pp. 934–940.
- Wiechetek, Linda & Jose Maria Arriola 2011: An Experiment of Use and Reuse of Verb Valency in Morphosyntactic Disambiguation and Machine Translation for Euskara and North Sámi. *NEALT Proceedings Series 2011*. Volum 14 [10].
- Wiechetek, Linda, Francis M. Tyers & Thomas Omma 2010: Shooting at flies in the dark: rule-based lexical selection for a minority language pair. *IceTAL'10 Proceedings of the 7th international conference on Advances in natural language processing*. Pages 418-429. Springer-Verlag Berlin, Heidelberg.

Andre referanser

- Beesley, Ken & Lauri Karttunen 2003: *Finite State Morphology*. CSLI Publications,
- Bick, Eckhard 1999: *The parsing system Palavras* Aarhus Univ. Press.
- Koehn, Philipp 2005: Europarl: A Parallel Corpus for Statistical Machine Translation. *The tenth Machine Translation Summit* page 79--86. Phuket, Thailand, AAMT. homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf
- Magga, Ole Henrik 2012: Lexicography and indigenous languages. *Euralex 2012 Proceedings*. www.euralex.org/elx_proceedings/Euralex2012/pp3-18%20Magga.pdf

Curriculum vitae – Trond Trosterud

Full name: Trond Trosterud
Date and Place of birth: Oslo, 30 August 1962

Current position:
Professor in Saami language technology at the University of Tromsø, leader of the Giellatekno Centre for Saami language technology

Private Address: Rundvannet 129. 9018 Tromsø; **Telephone** (work): 77644763

Education:

Dr.art. from Universitetet i Tromsø in 2004, with the dissertation *Homonymy in the Uralic Argument Agreement Paradigm*. Supervisors: Prof. Knut Tarald Taraldsen, (University of Tromsø) and Prof. Alho Alhoniemi (University of Turku).

Cand.philol. from Universitetet i Trondheim in 1990, with the subjects Linguistics (main subject, (hovudfag) (grade: 1.7 *laudabilis* [1.0: highest – 4.0 lowest]), Nordic linguistics (mellomfag), Finnish (grunnfag), History (grunnfag), and the minor subjects Mathematics, Statistics and Norwegian in a professional context.

M.A.Thesis (Hovudoppgåve) (grade: 1.6): *Binding Relations in two Finnmark Finnish Dialects. A Comparative Syntactic Study*. Supervisor: Prof. Lars Hellan, University of Trondheim.

Previous professional appointments:

Last 3 years: Supervisor for 5 different phD candidates within language technology and Saami sociolinguistics. One of them (Linda Wiechetek) is expected to deliver her dissertation this winter.

.

Previous professional appointments:

2006-present: Head of the Giellatekno research unit, University of Tromsø.

2001-2006: Worked as a researcher on projects for development of Saami language techno-logy, (http://giellatekno.uit.no/).

2000-2001: Worked for Lingsoft, Inc., in Helsinki, Finland, with two-level morphology and constraint grammar, participating in a team making the first grammar checker software for Norwegian (licensed by Microsoft for their Office suite).

1999: University lecturer, Department of Linguistics, University of Tromsø.

1999-2000: University lecturer, Department of Finnish, University of Tromsø.

1995-1997: Consultant for indigenous peoples' affairs, and secretary of the Indigenous Peoples' Committee. The Barents Secretariat, Kirkenes, Norway.

Other Activities:

2013- Vice-President of NEALT, the North European Association of Language Technology.

2013- Deputy member of the board of Gáldu, a resource centre for the rights of indigenous people.

2011- Member of the board of the Norwegian National Language Corpus Repository (Språkbanken).

2011- Member of the board of the Norwegian Language Council.

2007-2008: Member of the international committee on standardisation of Cornish, later as an arbiter in the same process.

2007- Member of the board of Kvensk institutt, the Norwegian institute for Kven affairs.

1997-2000: Norway's representative in ISO/IEC JTC1 WG2, WG3 Character set standardization.

1996-2000: Member of the Saami Committee for digitalisation and localisation

Research projects, funding and tangible outputs:

2006- Founded the Giellatekno language technology research centre, now together with the Divvun group a permanent centre at the university with 9 employers

- Received full-time funding from the Norwegian Research Council first for one researcher (Trosterud) 2001-, then two other researchers forming the Giellatekno research group 2006-, now fully permanently funded by the University of Tromsø.

- Initiated the Divvun project starting from 2004- with full-time funding by the Norwegian Sami parliament of up to six project workers, which are since 2011- fully permanently funded by the University of Tromsø.

The Giellatekno and Divvun projects have resulted in the following software tools:

- spellcheckers for North, Lule and South Saami, in use by all school children (9 years and older), all L2 learners, and most L1 users, since 2008.

Making a North Saami spell checker was a prerequisite for the establishment of the daily North Saami newspaper Ávvir, http://avvir.no
- Pedagogical programs for North and South Saami, with 350000 queries for the North Saami programs, over a 4-year period, for a language community of 18000 speakers.
- Electronic dictionaries for approx. 10 different languages, in daily use.

1995-1997: Initiated the *Corpus project for smaller Uralic languages*, funded by a Nordic grant, for 5 researchers.

List of publications

Scientific publications

70. Lene Antonsen, Ryan Johnson, Trond Trosterud and Heli Uibo 2013: Generating modular grammar exercises with finite-state transducers. In: Elena Volodina, Lars Borin, Hrafn Loftsson (eds.): *Proceedings from the 2nd workshop on NLP for computer-assisted language learning*. url: http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/64_Antonsen_etal.pdf

69. Sjur N. Moshagen, Tommi A. Pirinen, Trond Trosterud 2013: Building an open-source development infrastructure for language technology projects. In: Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen (eds.): *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16. pp. 343-352. url: http://www.ep.liu.se/ecp/085/031/ecp1385031.pdf

68. Ryan Johnson, Lene Antonsen, Trond Trosterud 2013: Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries. In: Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen (eds.): *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16. pp. 59-71. url: http://www.ep.liu.se/ecp/085/010/ecp1385010.pdf

66. Trosterud, Trond; Nystad, Berit Merete. 2012: A North Sami translator's mailing list seen as a key to minority language lexicography. In: *Euralex 2012 Proceedings: Euralex International Association for Lexicography* 2012 ISBN 978-82-303-2228-4. s. 250-256

65. Trosterud, Trond 2012: A restricted freedom of choice: Linguistic diversity in the digital landscape. *Nordlyd* 2012 ;Volum 39.(2) s. 89-104

63. Antonsen, Lene; Trosterud, Trond. 2011: Next to nothing – a cheap South Saami disambiguator. *NEALT Proceedings Series* 2011 ;Volum 14.(10) s. -

62. Trosterud, Trond 2010: Felles leksikalske ressursar for språkteknologi og leksikografi. *LexicoNordica* 2010 ;Volum 17. s. 211-223

61. Antonsen, Lene, Trond Trosterud and Linda Wiechetek 2010: Reusing Grammatical Resources for New Languages. *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10). European Language Resources Association 2010 ISBN 2-9517408-6-7. pp. 2782-2789.

60. Antonsen, Lene ja Trond Trosterud 2010: Manne dihtor galgá máhttit grammatihka? *Sámi dieđalaš áigečála* 1:3-28.

59. Antonsen, Lene, Ciprian-Virgil Gerstenberger, Sjur N. Moshagen og Trond Trosterud 2009: Ei intelligent elektronisk ordbok for samisk. *LexicoNordica 16*.

56. Lene Antonsen, Biret Ánne Bals Baal, Saara Huhmarniemi ja Trond Trosterud 2009: Dihtor ja giela válljenvejolašvuođat – gielalaš ja pedagogalaš čuolmmat. Sáhkavuoruin sáhkan. Sámegiela ja sámi girjjálašvuođa muhtin áigeguovdilis dutkanfáttát. *Dieđut* 1/2009 pp. 86-101.

55. Kevin Unhammer and Trond Trosterud 2009: Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. J.A. Pérez-Ortiz, F. Sánchez-Martínez, F.M. Tyers (eds.) *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, p. 35–42 Alacant, Spain, November 2009. http://hdl.handle.net/10045/12025.

54. Tyers, Francis M., Linda Wiechetek and Trond Trosterud 2009: Developing Prototypes for Machine Translation between Two Sami Languages. Proceedings of the 13th Annual Conference of the EAMT, pp. 120-127.

53. Antonsen, Lene, Saara Huhmarniemi and Trond Trosterud 2009: Interactive pedagogical programs based on constraint grammar. *Proceedings of the 17th Nordic Conference of Computational Linguistics*. Nealt Proceedings Series 4.

52. Muhirwe, Jackson and Trond Trosterud 2008: Finite State Solutions For Reduplication In Kinyarwanda Language. *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* pp. 73-80. http://www.aclweb.org/anthology-new/I/I08/I08-3013.pdf

51. Trosterud, Trond 2008: Verbh. En sydsamisk verbhandbok. *LexicoNordica* 15:347-354.

48. Trosterud, Trond og Linda Wiechetek 2007: Disambiguering av homonymi i nord- og lulesamisk. Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007. Doaimmahan Jussi Ylikoski ja Ante Aikio. *Mémoires de la Société Finno-Ougrienne* 253, pp. 401-422.

Trond Trosterud, samisk 1.2.14-31.1.17

		Totalkostnader					Søkes Forskningsrådet					Egenandel HSLF				
Personal- og indirekte kostnader	Omfang	Total	2014	2015	2016	2017	Total	2014	2015	2016	2017	Total	2014	2015	2016	2017
Gerstenberg, samlet 50%		459	92	190	97	80						459	92	190	97	80
Prosjektleder Trosterud	25 %	986	287	295	304	100	0					986	287	295	304	100
Postdoc	100 %	2064		1018	1046		1966		967	999		98		51	47	
PhD	100 %	2910	858	967	999	86	2910	858	967	999	86					
Studentstipend		200		100	100		200		100	100						
Tyers	10 %	381	92	101	103	85	0					381	92	101	103	85
Antonsen, samlet 50%		499	100	206	106	87						499	100	206	106	87
Sum lønn		7040	1337	2687	2658	358	5076	858	2034	2098	86	1964	479	653	560	272
Driftskostnader:																
Programmering		200		100	100		200		100	100						
Konferanser		100	25	25	25	25	100	25	25	25	25					
Samlet drift		300	25	125	125	25	300	25	125	125	25					
Samlet lønn/drift		7340	1362	2812	2783	383	5376	883	2159	2223	111					