Maskinoversetting fra nordsamisk, bra eller dårlig for sørsamisk?

Lene Antonsen
Giellatekno, Institutt for språk og kultur
UiT Norges arktiske universitet

Forskjellige formål

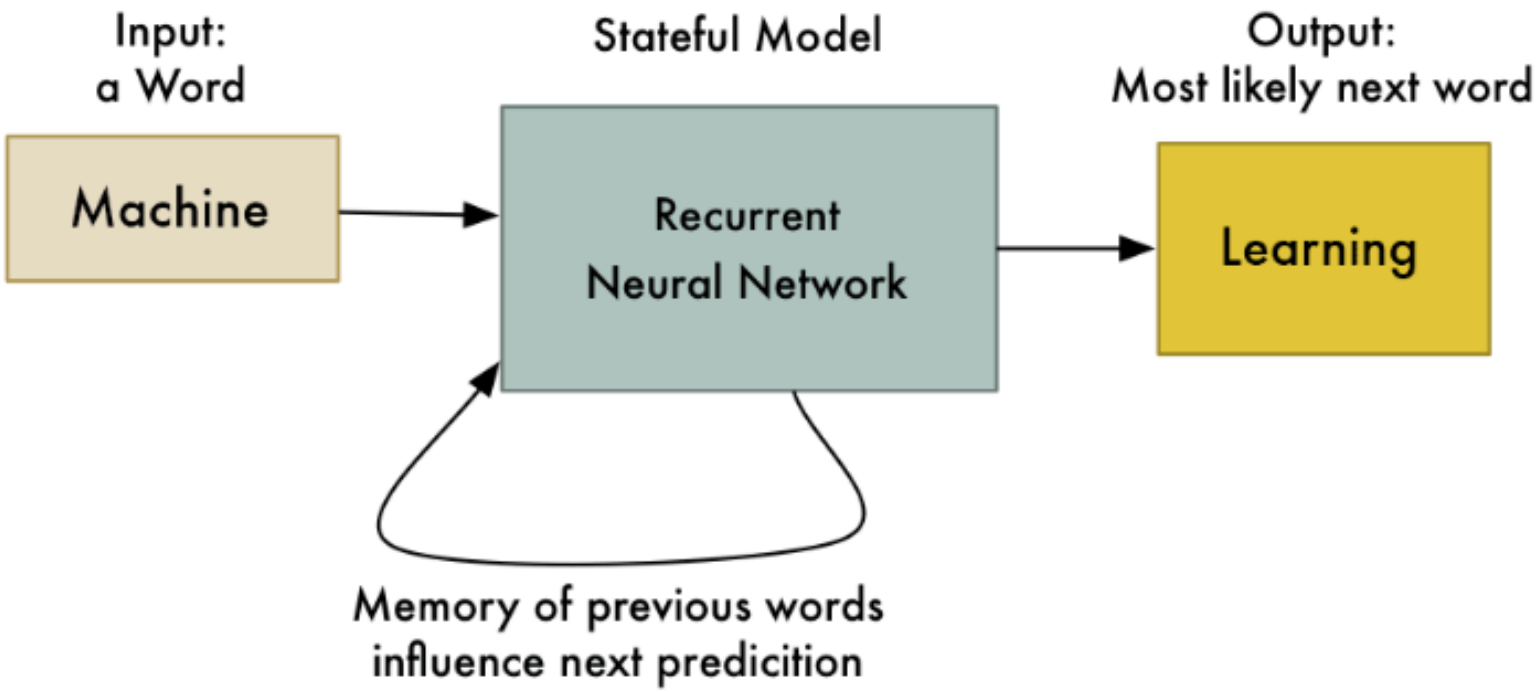
Maskinoversetting (MT)

- For forståelse (vi kan skrive på samisk, de som ikke forstår bruker MT)
- Oversetting for små domener: Menyer/lokalisering, manualer, værmelding...
- For posteditering (hjelpemiddel for oversetteren)

Statistisk maskinoversetting og nevrale nettverks maskinoversetting

- Bygger oversettinga ved å bruke statistiske metoder i et stort tospråklig tekstkorpus.
- Programmet 'ser' en eller flere ordformer, og bare deres **ytre form**, omd:
  mii čálisteimmet
  som den søker en sannsynlig oversetting til i korpuset
- Nevral maskinoversetting: Bruker et stor kunstig nevralt nettverk, og ved hjelp av deep learning kan programmet regne ut hva som er mest sannsynlig.

Nevral maskinoversetting



Samiske språk er veldig annerledes enn norsk og engelsk

Engelsk verb: walk – 4 forskjellige former:
walk, walks, walked, walking

Lulesamisk verb: vádtset – 56 forskjellige former:
vádtset, vádtse, vádtseba, vádtsebihtit, vádtsebihtte, vádtsem, vádtsema, vádtsemijn, vádtsemin, vádtsemis, vádtsep, vádtsi, vádtsin, vádtsis, vádtsisa, vádtsiska, vádtsit, vádtsiv, vádtson, vádtsop, vádtsu, vádtsuda, vádtsun, vádtsup, vádtsus, vádtsusa, vádtsuska, vádtsut, vádtsá, vádtsám, váttse, váttsedijn, váttsek, váttsij, váttsijda, váttsijga, váttsijma, váttsisj, váttsitja, váttsitjav, váttsitji, váttsitjibá, váttsitjihpit, váttsitjihppe, váttsitjin, váttsitjip, váttsá, váttsálulu, váttsáluluj, váttsálulujda, váttsálulujga, váttsálulujma, váttsálulun, váttsáluluv, váttsátjit, váttsáv

og så kommer ordavledninger ...
Dette krever enda større tospråklige korpus enn for engelsk og norsk.

Samisk maskinoversetting er regelbasert

Arbeid (fra nordsamisk)

- Eksperimenter 2008–2009: til lulesamisk
- Eksperimenter 2013: til sørsamisk
- Norges forskningsråd prosjekt 2014–2016:
  - til enaresamisk
    - også Kone Foundation og Finnish Cultural Foundation 2015 har støttet arbeidet
  - til lulesamisk
  - til sørsamisk
- til norsk (ikke egen finansiering)

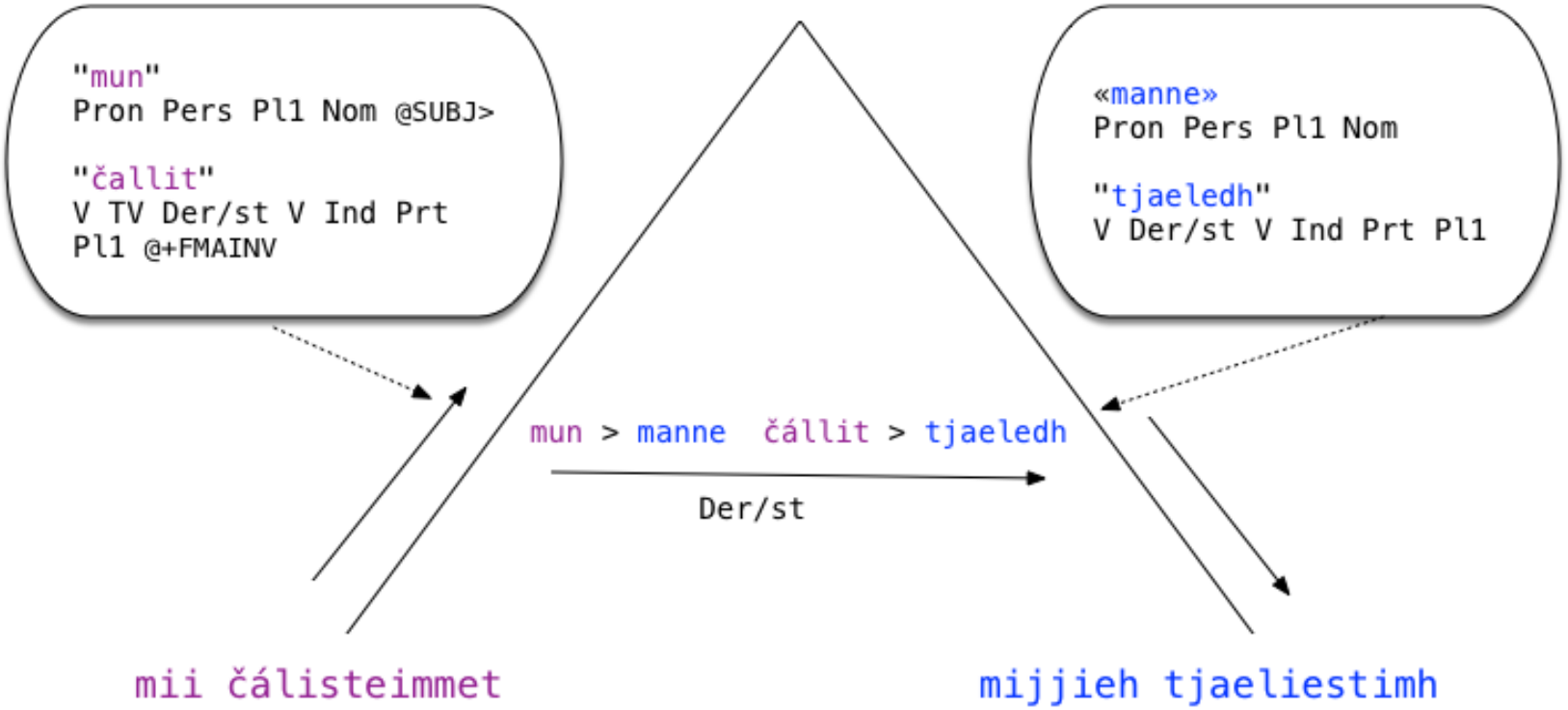
Nordsamisk-sørsamisk:
Francis Tyers, Kevin Unhammer, Trond Trosterud, Lene Antonsen, Maja Kappfjell, Anja Labj

Regelbasert oversetting

```
"<mii>"
    "mun" Pron Pers Pl1 Nom @SUBJ>
"<čálisteimmet>"
    "čállit" Ex/V TV Der/st V Ind Prt Pl1 @+FMAINV
```

- Grunnlaget er grunnformen pluss grammatiske tagger
- Programmet trenger regler: generelle regler og spesialregler (alt må beskrives eksplisitt)

Regelbasert oversetting



Velge riktig ord utfra kontekst

Ved flere mulige oversettelser: Systemet velger oversettelsen av et ord i henhold til regler, f.eks. "lohkat": lohkedh, jiehtedh, ryöknedh

Semantiske tagger hjelper i valg av oversetting:

nieida N **Sem/Hum** (human)
girji N **Sem/Txt** (text)
sámegiella N **Sem/Lang** (language)
gávpot N **Sem/Plc** (place)
...
(104 forskjellige semantiske tagger)

Fra nordsamisk til sørsamisk, noen eksempler

- Nords. har ett kasus, sørsamisk har to: **akkusativ og genitiv**
  - Analysen av nords. har tagger for akkusativ og genitiv
- Nords. har lokativ, sørsamisk har **elativ og inessiv**
  - I analysen av nords. legges det til elativ- og inessiv-tagger
- Nords. habitiv-advl. er lokativ (jeg har), sørsamisk er **genitiv**
  - I analysen av nords. legges til <hab> tagg, som utløser endring fra lokativ til genitiv i MT-programmet
- Nords. har for det meste SVO, sørsamisk har **SOV**
  - I MT-programmet flyttes hovedverbet
- Sørsamisk har ofte ikke **lea-verbet**
  - I MT-programmet fjernes *lea* i en del tilfeller

Regelbasert oversetting

```
"<Mii>"
    "mun" Pron Pers Pl1 Nom @SUBJ>
"<lohkagoahtit>"
    "lohkat" Ex/V TV Der/InchL V Ind Prs Pl1 @+FMAINV
"<girjji>"
    "girji" N Sem/Txt Sg Acc @<OBJ
```

Flytt hovedverbet (FMAINV) forbi objektet (OBJ):

- Mii lohkagoahtit girjji → Mijjieh gærjam lohkegåetiejibie

Alternativ: Gjør om Der/InchL: aelkedh+morfologi verb+Inf

- Mii lohkagoahtit girjji → Mijjieh aelkiejibie gærjam lohkedh

Ordlista i MT-programmet må utvides

```
<e><p><l>bargoáigi<s n="n"/></l><r>barkoetïjje<s n="n"/></r></p></e>
<e><p><l>bargu<s n="n"/></l><r>barkoe<s n="n"/></r></p></e>
<e><p><l>barta<s n="n"/></l><r>hahtjoe<s n="n"/></r></p></e>
<e><p><l>basadanlatnja<s n="n"/></l><r>laavkoe<s n="n"/></r></p></e>
<e><p><l>bassi<s n="n"/><s n="nomag"/></l><r>bïssije<s n="n"/></r></p></e>
<e><p><l>bassi<s n="n"/><s n="sem_time"/></l><r>bissiebiejjie<s n="n"/></r></
<e><p><l>baste<s n="n"/></l><r>buste<s n="n"/></r></p></e>
<e><p><l>bavssa<s n="n"/></l><r>bangseme<s n="n"/></r></p></e>
<e><p><l>beaggin<s n="n"/></l><r>såaltje<s n="n"/></r></p></e>

<e><p><l>Maj<s n="np"/></l><r>Maj<s n="np"/></r></p></e>
<e><p><l>Majabritt<s n="np"/></l><r>Majabritt<s n="np"/></r></p></e>
<e><p><l>Majavatn<s n="np"/></l><r>Maajehjaevrie<s n="np"/></r></p></e>
```

52 800 ordpar

Bare 6400 som ikke er navn!!

Bygging av ord 1

Hvis ordparet ikke finnes i ordlista inne i MT-programmet, så bygges det etter nordsamisk mønster (hvis delene finnes)

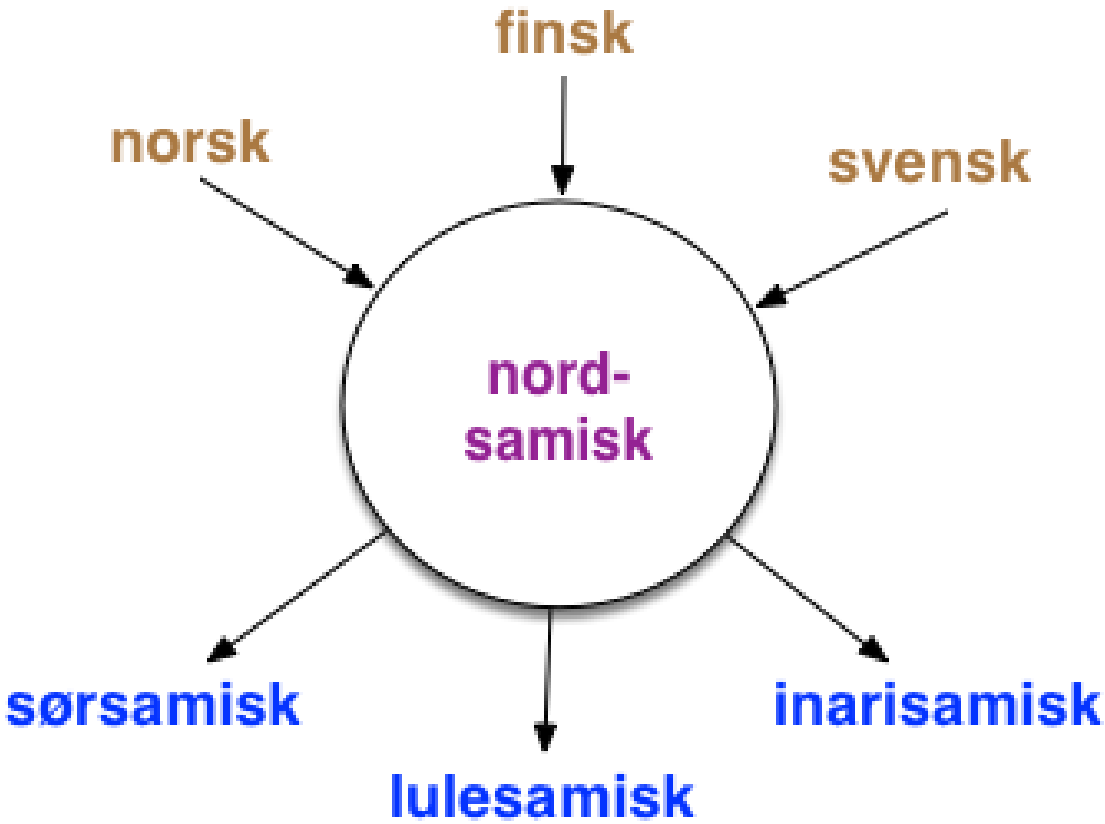
- lakseoppdrettsindustri
- luossabiebmanindustriija
- luossa-biebman-industriija
- loese-bïepmehtimmie-industrije
- **loesebïepmehtimmieindustrije**

Bygging av ord 2

- klesdesigner
- bivttashábmejeaddji
- bivttas-hábmejeaddji $<$ hábmet
- gaarvoe-hammoedæjja $<$ hammoedidh
- **gaarvoehammoedæjja**

Nordsamisk som nav

Manuell oversetting til nordsamisk, maskinoversetting fra nordsamisk:



Nordsamisk som ressurs

MT fra nordsamisk

Dáppe galgá okta dievdu jávkkihan eret, go doaktárat galge sus váldit varraiskosiid, maŋŋel go politiijat ledje balahan su vuoddján gárrenoaivvis biillain baddjelluotta. Govven: Ellen Inga Kristine Hætta (vuorkkágovva)

Dievdu báhtarii go galggai varraiskosiid addit

Politiijat bidje guokte dievddu ambulánsabargiid háldui váldin dihte varraiskosiid, maŋŋel go leigga balahuvvon gárrenoaivvis vuoddján biillain. Nubbi dievdu ii doahttalan doaktáriid ii ge ambulánsabargiid, ja guđii dearvvašvuođaguovddaža Guovdageainnus nu eai beassan váldit varraiskosiid.

Čállán: Nils
Johan Var
Journalista

12:50 Golg

Lávvorda
biila lea v
lulábeald

Olbmot, g
gávdne v

Gávdne gu

Oarje-Finn
muitala pol
guokte diev

Goappešat

Báhtarii

Nimetön 4 — Muokattu

Mannen flyktet da han skulle gi blodprøvene

Politiene satte to menn ambulansearbeidere til besittelsen ta for blodprøvene, etter at de to hadde balahuvvon alkoholpåvirket kjørt med en bil. Annen mannen respekterte ikke leger han ikke også ambulansearbeidere, og forlot dearvvašvuođaguovddaža i Kautokeino slik de fikk ikke ta blodprøver.

Nimetön 5 — Muokattu

Mies pakeni kun täytyi verikokeita antaa

Poliisit paivat kaksi miestä ambulanssityöntekijöitä mielivaltaan kaappaamiseksi takia verikokeiden, jälkeen kun olivat balahuvvon gárrenoaivvis ajajani auton. Toinen mies ei doahttalan lääkäreiden eikin ambulanssityöntekijöitä, ja jätti terveydenkeskusta Koutokeinossa näin eivät päässeet kaapata verikokeita.

Nimetön 2 — Muokattu

Alma baatari gosse edtji virrepryövenassh vedtedh

Polijsh göökte almah bïejin ambulánsabargiid ritnemasse vaeltemen diehre virrepryövenassi, mænngan gosse ligan beavnasovveme gárrenoaivvis vuajajamme bijline. Mubpie alma idtji dåakterh ussjedh ij aaj ambulánsabargiid, jïh laehpieji dearvvašvuođaguovddaža Goevtegeajnosne dan idtjin åadtjoeh virrepryövenassh vaeltedh.

Nimetön 3 — Muokattu

Ålmåj báhtarij gå galgaj varraåtsålvisájt vaddet

Politijja biedjin guovtev ålmmåv ambulánsabargiid hállduj válldema diehti varraåtsålvisáj, maŋŋela gå lijga baláduvvam gárrenoaivvis vuoddjám bijlajn. Nubbe ålmåj ij doahttalan dåktárij ij ga ambulánsabargiid, ja guottij dearvvašvuođaguovddaža Guovddagæjnon nav ettjin besa válldet varraåtsålvisájt.

Nimetön — Muokattu

Almai patârij ko koolgâi vorrâiskosijd adeliđ

Pooliseh pieijii kyehti almaa ambulanspargei haaldun väldiđ tiet vorrâiskosij, maŋa ko láin poollâđ gárrenoaivvis vyeijem autoin. Nubbe almai ij doahttalan tuáhtárij ij-uv ambulanspargeid, já kuođij tiervâsvuođâkuávdáá Kuovdâkiäinust nuuvt iä peessâm väldiđ vorrâiskosijd.

Våre MT-språkpar

Den nordsamiske analysatoren er best utbygd og passer derfor best som **fra**-språk. Dessuten publiseres det mest på nordsamisk.
Dagens situasjon:

- Fungerer best
  - Fra nordsamisk til inarisamisk
  - Fra nordsamisk til lulesamisk
  - Fra nordsamisk til norsk (for forståelse)
    Også i bruk i akademia
- Store syntaktiske utfordringer
  - Fra nordsamisk til sørsamisk
- Ord-til-ord oversettelse (Også i bruk i akademia)
  - Fra inarisamisk til nordsamisk
  - Fra lulesamisk til nordsamisk
  - Fra sørsamisk til nordsamisk

Mye kan gå feil

- Feil språk i kilde-språket
- Feil analyse av kilde-språket
- Oversetting mangler i ordlista
- Oversetting i ordlista passer ikke til konteksten
- Feil ordstilling
- Feil bøyningsform

Hvordan forbedre programmet?

- Feil språk i kilde-språket ← bruke Divvun-programmet
- Feil analyse av kilde-språket ← forbedre analysatoren
- Oversetting mangler i ordlista ← legge til flere ord i ordlista
- Oversetting i ordlista passer ikke til konteksten ← legge til ord og regel
- Feil ordstilling ← lage regel
- Feil bøyningsform ← forbedre regel

Vi må samarbeide om å forbedre programmet

Saemien Gïelegaaltije




oversettere, studenter, UiT

Varianter av sørsamisk

Bohten ikte. => Jåååktan båetiejim.

Varianter av sørsamisk

Bohten ikte. => Jååktan böötim/båetiejim.

Varianter av sørsamisk

Man kan lage to varianter av MT-programmet:

Bohten ikte. => Jååktan båetiejim.
Bohten ikte. => Jååktan böötim.

Kildespråket påvirker, samisk mellom øst og vest

Genitiv attributt eller komplement?

Fylkesmannen i Nordland var prosjektleder.
Fylhkemaennie Nordlaantesne ...
Via nordsamisk med MT:
Nordlándda fylkamánne lei prošeaktajođiheaddji.
Nordlaanten fylhkemaennie lij prosjektejuhtiehtæjja.

Rapporten fra Telemarksforskningen viser ...
Reektehtse Telemarksforskningeste vuesehte ...
Via nordsamisk med MT:
Telemarkforskningen raporta čájeha ...
Telemarkforskningen reektehts vuesehte ...

Via nordsamisk, eksempel

Barnehagen ledes av samiskspråklige pedagogisk personale.
Maanagïerte stuvresåvva saemien pedagogeles barkijijstie.

Via nordsamisk med MT:
Mánáidgárddi jođihit sámi pedagogalaš bargit.
Maanagïertem saemien pedagogen barkijh mietiedieh.

Oversettingsprogrammet

`http://gtweb.uit.no/mt-testing/`

Hva vil dere ha for sørsamisk?

Maskinoversetting

- Ingen
- Fra nordsamisk for posteditering, hjelpemiddel for oversettere et bestemt domene
- Fra nordsamisk, hjelpemiddel for oversettere generelt
- For sørsamisk til nordsamisk, for forståelse
- For forståelse, til norsk
- Annet?