Språkteknologi for meänkieli

Trond Trosterud, Giellatekno, UiT

6.12. 2020

Eg har i haust bygd ein språkmodell for meänkieli, i form av ein endeleg tilstandstransduer (finite state transducer), med programmeringsspråka *lexc* og *twolc*[1]. Kjeldekoden er publisert på github, på adressa `http://github.com/giellalt/lang-fit`. Kjeldekoden inngår som ein del av infrastrukturen for språkteknologi for ulike minoritetsspråk utarbeidd av forskingstruppene *Giellatekno* og *Divvun* ved UiT Noregs arktiske universitet. Dene rapporten går først gjennom datamodellen for meänkieli og ser deretter nærare på kva tilstand språkmodellen er i. Til slutt kjem nokre råd for framtidig arbeid med meänkieli.

1 Ein maskinlesbar språkmodell for meänkieli

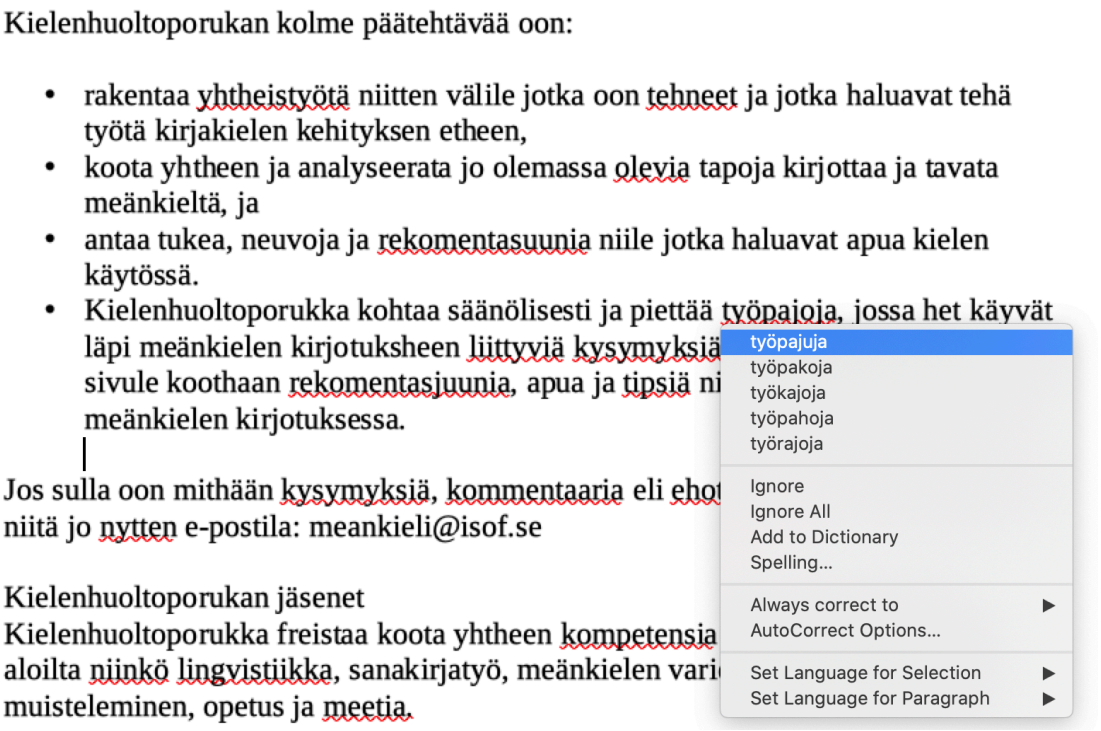
Språkmodellen for meänkieli er ein tovegs modell som både er i stand til å analyseere ordformer (seie at *kielen* er eintal genitiv av *kieli*) og å generere ordførmer frå lemma og grammatisk spesifikasjon (gje alle kasusformene av ordet *kieli*). I og med at språkmodellen er ein del av denne infrastrukturen, er det mogleg å bruke han i ulike praktiske program. Til no er språkmodellen i bruk i tre ulike samanhengar:

1. Som retteprogram for ulike operativsystem og dataprogram
2. Som program for å analysere tekst på meänkieli
3. Som program for generering av bøyingsparadigmer for meänkieli

For å bruke retteprogrammet for meänkieli kan brukaren gå til Divvun-gruppa si nedlastingsside `http://divvun.no`, og laste ned installeringsprogrammet *Divvun Installer*. Programmet blir installert på maskina på vanleg måte. Installeraren skil mellom "Divvunspråk" og "Alle språk", meänkieli ligg under lista "Alle språk", under namnet *Tornedalen Finnish* (som er namnet språket har i den relevante ISO-standarden), og ved å velje tornedalsfinsk vil brukaren få installert retteprogrammet på maskina si. I tillegg vil installeringsprogrammet også sørge for at retteprogrammet blir automatisk oppdatert. På Windows vil installeringsprogrammet installere retteprogrammet for Microsoft Word, mens det på Macintosh vil installere retteprogrammet på systemnivå (men ikkje for Microsoft Word). Installeringsprogramemt er under utvikling, og det vil installere retteprogrammet i fleire program etter kvart.

Figur 1 viser eit døme på bruken av retteprogrammet, her for LibreOffice. Figuren viser at programmet er langt frå ferdig, det er mange korrekte ord som ikkje blir kjent att, såkalla *falske positive*. Sjå neste avsnitt for ei evaluering av språkmodellen.

[1] Desse programmeringsspråka er dokumentert i `http://fsmbook.com`.



Figur 1: Retteprogram for meänkieli i LibreOffice. Programmet kjenner ikkje att *työpajoja*, og gjer framlegg om *työpajuja*

Språkmodellen er også sett inn i to interaktive nettsider, ei for analyse av tekst, jf. figur 2, som viser ei av setningane frå figur 1. På denne nettsida er det mogleg å lime inn tekst og få ei morfologisk analyse (framleis uavhengig av kontekst). For å unngå dataangrep har vi sett ei grense på 250 ord i dette grensesnittet, med å bruke programmet på eiga datamaskin er det mogleg å analysere så mykje tekst ein vil. Det er også mogleg å generere bøyingsparadigmer for dei ulike opne ordklassene, jf. `http://giellatekno.uit.no/cgi/p-fit.nob.html`.

Sjølve språkmodellen er tilgjengeleg som open kjeldekode, i github[2]. Meänkieli er ikkje det einaste språket vi har utarbeidd språkteknologi for ved UiT, vi har fullskalaspråkmodellar for over 30 språk, ein tredel av dei er i aktiv bruk[3].

[2]Filene med dei grammatiske modellane ligg på `https://github.com/giellalt/lang-fit/tree/develop/src/fst`, dei relevante filene er `phonology.twolc` og katalogane `stems` og `affixes`. Hovudsida for meänkieli-prosjektet er `https://github.com/giellalt/lang-fit`

[3]Ei liste over dei ulike språka, ei vurdering av kvaliteten på dei ulike språkmodellane og kva lisens dei er tilgjengeleg under ligg på `https://github.com/divvun/registry`, under **Languages** eit stykke ned på sida. Det er lenkjer til kjeldekoden i kvart tilfelle

```
koota   koota+V+Inf
koota   koota+V+Pass+Ind+Prs+ConNeg

yhtheen yksi+Pron+Indef+Ill
yhtheen yksi+Num+Sg+Ill
yhtheen yhtheen+Adv

ja      ja+CC

analyseerata    analyseerata+V+Inf
analyseerata    analyseerata+V+Pass+Ind+Prs+ConNeg

jo      jo+Adv

olemassa        olla+V+InfMa+Ine
olemassa        olla+V+InfMa+Ine

olevia  olevia  +?

tapoja  tapa+N+Pl+Par

kirjottaa       kirjottaa+V+Inf

ja      ja+CC

tavata  tavata+V+Inf
tavata  tavata+V+Pass+Ind+Prs+ConNeg

meänkieltä      meänkieli+N+Sg+Par
meänkieltä      meänkieli+N+Sg+Par

.       .+CLB
```

Figur 2: Tekstanalyse på `http://giellatekno.uit.no/cgi/d-fit.nob.html`

2 Tilstanden for språkmodellen for meänkieli

Språkmodellen eg bygd opp av tre komponentar: lemmalista til ordboka *Meänkieli - ruotti sanakirja*, skrive for Academia Tornedaliensis av Linnea Nylund, Märta Nylund, Birgitta Rantatalo og Erling Wande, ein morfologisk og ein morfofonologisk komponent som eg har skrive, basert framfor alt på Kenttä og Pohjanens *Meänkielen kramatiikki* frå 1996. Ein tisvarande modell vart laga for kvensk som eit samarbeid mellom UiT og Kainun institutti, jf. Trosterud, S. m.fl. 2017. Denne artikkelen kan også lesast som ei innføring i den lingvistiske modellen for meänkieli[4].

Språkmodellen er langt frå ferdig. Leksikonet inneheld 25000 lemma (av dei 15000 substantiv, 5000 verb og 2000 adjektiv) frå ordboka. Rundt 51 % av dei opne ordklassene er ikkje klassifisert, og heile leksikon bør eigentleg bli gått gjennom. I tillegg inneheld leksikon ca. 32000 namn (stadnamn, personnamn), tatt frå dei generelle språkressursane ved UiT. Det er også hol i den morfologiske

[4]Trosterud, S. m.fl. 2017: A morphological analyser for Kven. Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages. `https://www.aclweb.org/anthology/W17-0608/`.

komponenten: Delar av dei infinitte formene er ikkje komplett, det same gjeld possessivsuffiksa. Derivasjonsmorfologien er framleis ikkje ein del av grammatikken, bortset frå avleiing av verb til substantiv med suffikset *-minen*. Delar av pronomensystemet er også uferdig.

Etter å ha arbeidd ein månad med administrative tekstar fekk eg dekningsgraden for eit korpus på 40000 ord bestående av administrativ tekst opp frå 72 % til 82 %. Deretter skifta eg til eit nytt korpus, *Poppimysiikkiä Vittulasta*, der eg fekk ein dekningsgrad på 77 % på første forsøk. Dette er godt nok til å vere nyttig for leksikografisk arbeid, t.d. for å kartlegge bruk av substantiv og verb hos Mikael Niemi. For ein god stavekontroll er dette ikkje godt nok, då bør dekningsprosenten ligge godt over 95 %, gjerne over 97 %. For å oppnå det er det fleire ting som må til.

Å få dekningsgraden opp til 90 % går erfaringsmessig relativt fort, på under eit halvt årsverk, det er nok å komplettere morfologien, rette leksikonet og legge inn klasser av ord som oppfører seg likt. Resultatet vil framleis vere eit ordretteprogram som har mellom ein og to falske alarmar i kvar setning. Å auke dekningsgraden frå 90 % til 95 % vil ta lenger tid enn arbeidet fram til 90 %, og deretter går det berre saktare og saktare å forbetre resultatet. Kvart ord som blir lagt til og kvar feil som blir retta vil berre ha minimal innverknad på dekningsgraden, likevel er det dette arbeidet som må gjerast for å få eit analyseprogram ein kan stole på.

Den morfofonologiske komponenten i analyseprogrammet er ikkje komplett, det er framleis morfofonologiske prosessar som ikkje er med. Kasus illativ er t.d. heller ikkje på plass for alle stammeklassar, og det finst fleire substantivstammetypar som framleis ikkje er gjort greie for. Deretter står det att eit arbeid med å formalisere den delen av grammatikken som ikkje står i grammatikkbøkene. Det er også fleire normative spørsmål som er opne, spørmsål som må løysast eksplisistt fordi analyseprogrammet treng å formalisere det.

Språkmodellen treng også eit disambigueringsprogram, for å skilje mellom t.d. verbet og substantivet *tulen*, men også homonymi av typen *net tuleva* vs. *tuleva vuosi*. Når dette er sagt er det viktig å ha i mente at språkmodellen allereie no er god nok til å vere nyttig.

3 Råd for framtidig arbeid med meänkieli

Retteprogram og tilsvarande program er ekstremt kraftige, ja eg vil nesten si uunnverlege verkty i språknormeringsarbeidet. Med eit retteprogram vil dei normative organa for meänkieli vere i stand til å få gjennomslag for normeringa av språket, utan eit slikt program blir det langt vanskelegare. Tilsvarande vil det for ein ny generalsjon av språkbrukarar vere slik at eit godt retteprogram vil kunne gje dei den tryggheita dei treng for å kunne bruke meänkieli skriftleg.

Meänkieli har status som minoritetsspråk i Sverige. Det inneber at språket vil bli brukt skriftleg i offentlege samanhengar, slik det til ein viss grad blir allereie. For å fungere i desse samanhengane treng meänkieli retteprogram. Eg har følgjande framlegg til dei som arbeider med planlegginga av revitalisering og normering av meänkieli:

- På kort sikt lagar Isof eller andre relevante organ ei arbeidsgruppe som kan forbetre språk-

modellen for meänkieli til eit nivå som gjer programmet bra nok til å bli tatt i bruk

- På lengre sikt bør normeringsarbeid og språkteknologi skje på permanent basis.

Rollefordelinga for arbeid med språkteknologi for meänkieli bør vere følgjande: Universitetet i Tromsø kan ikkje vere den som styrer arbeidet med meänkieli, det må skje frå Sverige. Derimot kan arbeidet med meänkieli bruke infrastrukturen vi har bygd opp for språkteknologi for minoritetsspråk på Github, og det kan dra nytte av arbeidet vi gjer for å integrere desse språkmodellane i Apple, Google og Microsoft sine produkt.

Vi ved UiT har erfaring både med å tilpasse infrastrukturen og med og lage språkmodellar. Det er på bakgrunn av dette vi har vore i stand til å lage språkmodell og retteprogram for meänkieli. Vår innsats i dette arbeidet kan i framtida vere: Vi kan også i framtida sørge for at språkmodellane for meänkieli kan bli brukt som retteprogram og andre praktiske program. Vi kan også bidra med råd når det gjeld vidareutviklinga av modellen, og vi kan gje opplæring i det språkteknologiske arbeidet. Alt dette kan vi gjere som del av det ordinære arbeidet vårt, utan separat finansiering.

I og med at denne språkteknologien er grammatikkbasert (kunnskapsbasert) har ikkje tekstinnsamling ei like sentral rolle som det ville ha hatt for metodar som bruker maskinlæring. Noko anna er heller ikkje mogleg: Eksisterande tekstar på meänkieli inneheld ofte uassimilerte lån både frå svensk og finsk, og dei inneheld former som avvik frå normalen Isof går inn for. Ved UiT har vi samla inn i underkant av 100000 ord, halvparten har vi funne på offisielle nettsider og halvparten har vi fått frå Bengt Pohjanen. Samanlikna med korpussamlingane vi har for samiske språk (nordsamisk 35 mill, enaresamisk 1,7 mill, sørsamisk 1,5 mill. og lulesamisk 1,3 mill ord) er dette lite, men i denne fasen av arbeidet er det likevel nok, og hovudvekta av arbeidet bør gå til arbeidet med språkmodellen heller enn til å samle inn tekst.

Det bør for Isof vere mogleg å til ein viss grad arbeide med språkmodellen innafor dei eksisterande økonomiske rammene som finst for meänkieli. Det å arbeide med språkmodellen er i praksis å arbeide med normering av leksikon og grammatikk for mänkieli, noko som nettopp er kjerna i språknormeringsarbeidet. Det å delta på opplæring i dette arbeidet i regi av UiT bør det også vere mogleg å få gjennomført innafor eksisterande økonomiske rammer. For å vere i stand til å arbeide vil meänkieli-lingvistane på Isof ha grunnleggjande kjennskap til arbeide på unix-kommandolinja. Dette er opplæring som vi på UiT for så vidt kan gje, men det vil vere meir effektivt viss datakyndig personell på Isof kan gje denne opplæringa, slik at vi på UiT kan bruke tida vår på det berre vi kan[5].

Det er med andre ord mogleg å gjere ein god del språkteknologisk arbeid for meänkieli også utan særskild finansiering, berre med å utnytte det som allereie er løyva til Isof i Sverige og til samisk språkteknologi i Noreg. For å få ordretteprogrammet for meänkieli opp til eit godt nivå trengst det likevel ein arbeidsinnsats tilsvarande fleire årsverk. Eigentleg bör denne finansieringa kome som ordinær løyving over det svenske statsbudsjettet, på same måte som det gjer i Noreg. Sverige har ratifisert Europarådet sin konvensjon for minoritetsspråk, og slått fast at meänkieli skal vere eitt av desse språka. Det å finansiere arbeidet med normative verkty som ordretteprogram er ein logisk konsekvens av denne ratifiseringa. Mens vi ventar på ei slik finansiering er det fleire ting som

[5]Eit oversyn over det lingvistane bør vite om unix er `https://giellalt.uit.no/tools/docu-unix-nno.html`.

kan gjerast. For det første kan vi arbeide innafor den eksisterande finansieringa. Det viktigaste er allereie oppnådd: Det finst eit fungerande retteprogram for meänkieli som alle kan laste ned og bruke. Det å løyve pengar til dette arbeidet er med andre ord ikkje risikabelt: Vi kan allereie no vise at programmet fungerer. Pengar til dette arbeidet vil ikkje gå til prøving og feiling, men til konkret og kvantifiserbar forbetring av programmet.

Isof bør legge konkrete planar for finansiering av arbeidet på både kort og lang sikt. På kort sikt gjeld det å identifisere middel til tidsavgrensa finansiering. Ei slik pengekjelde er faktisk Isof sjølv[6]. Det kan godt hende ein mogleg modell er eit trepartsamarbeid: UiT gjev teknisk og språkteknologisk støtte, Isof bidreg med normeringsarbeid, og ein tredjepart som søker på prosjektmiddel 21.1.2021 kan gjere det konkrete arbeidet. Tilsvarande kan det også finnast andre finansieringskjelder. Ei slik kjelde var utlysinga til Vetenskapsrådet i fjor, der gjekk søknaden for språkteknologi for minoritetsspråk dessverre ikkje gjennom. Eit meir konkret grunnlagsarbeid, som dette, kan likevel gjere det meir sannsynleg å nå opp i neste runde.

Ein viktig samarbeidspartner vil vere det pågåande bibelomsetjingsprosjektet for meänkieli. Dei som omset Bibelen tar normeringsarbeidet alvorleg, og vil gjerne skrive orda dei skriv konsistent gjennom heile teksten. Dette prosjektet er sannsynlegvis også det kvantitativt sett største pågåande prosjektet for skriftkultur for meänkieli akkurat no.

Den naturlege rolla for Isof i dette arbeidet vil vere administrativ og delvis også substansiell: Administrativ på den måten at sekretæren for meänkieliarbeidet på Isof bør ha oversikt over landskapet, vere innsett i situasjonen for språkteknologi for meänkieli og dra i dei trådane som trengst for å få arbeidet til å gå framover. Ambisjonsnivået for arbeidet er eigentleg allereide til stade: Sverige har fem nasjonale minoritetsspråk, der to av dei er makronamn for til saman ca. 10 ulike skriftspråk, og må dermed også gje desse ca. 15 språka vilkår for å kunne fungere som skriftspråk. Det å lage ordretteprogram for desse språka er ein føresetnad for å kunne fungere som nasjonale minoritetsspråk, og infrastrukturen og arbeidsmåten skissert i denne rapporten er ein måte å få dette arbeidet gjort på.

[6] Jf. utlysinga `https://www.isof.se/sprak/minoritetssprak/bidrag-till-minoritetsspraken/soka-bidrag.html`.