Samisk språkteknologi ved Universitetet i Tromsø

Disambiguering og syntaktisk analyse(1) ◊ Samisk stavekontroll(2)

(1): Lene Antonsen, Saara Huhmarniemi, Marit Julien, Ilona Kivinen, Trond Trosterud, Linda Wiechetek http://giellatekno.uit.no
2): Børre Gaup, Sjur Moshagen, Thomas Omma, Maaren Palismaa, Tomi Pieski (Sametinget i Noreg) http://divvun.no

SÁMEDIGGI SAMETINGET

UNIVERSITETET I TROMSØ

R e s u l t a t — s a m i s k

Grammatisk analyse

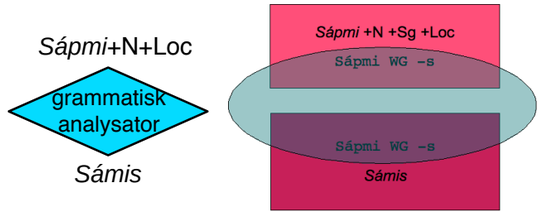
Vi har laga ein grammatisk analysator for nord- og lulesamisk tekst (ordbøying og -avleiing, lydovergangar), og ein setningsanalysator for nord- og lulesamisk (den nordsamiske har presisjon = 97,5 % og recall = 94,4 %).

Vi har laga eit syntaktisk annotert korpus for nordsamisk, som vi har gjort tilgjengeleg for språkforskarar over internett.

Stavekontroll

Den grammatiske analysatoren er grunnstammen i Sametinget sin stavekontroll for nord- og lulesamisk, som blir publisert i slutten av 2007 (betaversjonen ferdig i februar 2007).

Programmet inneheld også eit orddeleprogram.



Tekst frå den samiske avisa *Min Áigi*, i det samiske retteprogrammet (korrekte former: čeabehii, gávnnahit, boađášii)

Ii juohke bohcco xeabehii
GPS-čeabetbáttiid dahje dan teknihka maid gavnnahit buoremussan ii eisege leat ulbmil bidjat juohke bohccui. – Bálgosiid siste leat earálágan vuogádagat ja várra dat boađašii leat siiddaid anus.

A n d r e s p r å k

Grønlandsk stavekontroll

Utan morfologiske automatar er det umogleg å analysere grønlandsk, eller lage stavekontrollar: dei fleste orda på grønlandsk tilsvarer ei heil setning på norsk. I oktober 2006 vart det første korrekturprogrammet for grønlandsk lansert, basert på infrastrukturen til dei samiske analysatorane.

Ordbøker for komi og kvensk

Tida då ordboks- og språkteknologiprosjekt var uavhengige av kvarandre er over.
Med å koble ordbok og grammatikk saman får vi eit kraftig analyseverkty.
Med utgangspunkt i arbeidet vårt med samisk terminologi og analyse har vi gjort ei komi-ordbok om til ein integrert ordbok og analysator, og samtidig til ein modell for arbeid med maskinomsetjing i framtida. Vi har også laga eit oppsett for arbeid med kvensk leksikografi.

Færøysk analysator

Vi har utarbeidd ein betaversjon for ein færøysk analysator, som kan danne utgangspunkt for den færøyske delen av eit dansk–færøysk maskinomsetjingsprogram.

Oqaasileriffik

Ataasinngorpat 16.10.2006. KUKKUNIIAAT siulleq aaneqarsinnaanngorluni piariissaaq. Aallerneq akeqanngilaq. Uani aallersoqarsinnaavoq.

Mandag den 16.10.2006 er version 1.0 af KUKKUNIIAAT (den grønlandske stavekontrol og orddeler til MSOffice) klar til download. Den er gratis. Tryk her

```
"<Ataasinngorpat>"
        "ataasinngor" V Int 3Pl @PRED
        "ataasinngor" V Ind 2Sg 3Sg @PRED
        "ataaseq" NNGUR V Con 3Sg @PRED
        "ataaseq" NNGUR V Int 3Pl @PRED
        "ataasinngor" V Con 3Sg @PRED
        "ataaseq" NNGUR V Ind 2Sg 3Sg @PRED
"<16.10.2006>"
        "16.10.2006" Num Abs @SUBJ
"<.>"
        "." CLB <<<
"<Kukkuniiaat>"
        "kukkuniiaat" N Abs Sg @SUBJ
        "Kukkuniiaat" N Abs Sg @SUBJ
        "Kukkuniiaat" N Abs Sg @OBJ
        "kukkuniiaat" N Abs Sg @OBJ
"<siulleq>"
        "siulleq" N Abs Sg @SUBJ
        "siulleq" N Abs Sg @OBJ
"<aaneqarsinnaanngorluni>"
        "aaneqarsinnaanngorluni" ?
"<piariissaaq>"
        "piareer" SSA V Ind 3Sg @PRED
        "pia" RIISSA V Ind 3Sg @PRED
"<.>"
        "." CLB <<<
"<Aallerneq>"
        "aaller" NIQ N Abs Sg @SUBJ
        "aallerneq" N Abs Sg @SUBJ
        "aallerneq" N Abs Sg @OBJ
        "aaller" NIQ N Abs Sg @OBJ
"<akeqanngilaq>"
        "akik" QAR NNGIT V Ind 3Sg @PRED
        "aki" QAR NNGIT V Ind 3Sg @PRED
"<.>"
        "." CLB <<<
"<Uani>"
        "uv" Adv Lok @ADVL
"<aallersoqarsinnaavoq>"
        "aallersoq" QAR SINNAA V Ind 3Sg @PRED
        "aaller" TUQ QAR SINNAA V Ind 3Sg @PRED
"<.>"
        "." CLB <<<
```

A n a l y s e m e t o d a r

Vi har brukt endelege tilstandsautomatar for morfologisk analyse. Til morfofonologien bruker vi tonivåmorfologi. Analysatorane våre er kompliert med kompilatorar frå Xerox, (*lexc*, *twolc* og *xfst*, jf. *http://www.fsmbook.com/*).

Til syntaktisk analyse har vi brukt ein grammatisk basert syntaktisk analysator (føringsgrammatikk), *vislcg* frå Syddansk universitet, jf. *http://visl.sdu.dk*.

Automatisk analyse av 1. Mosebok, v. 1–2:

```
"<Álggus>"
        "álgu" N Sg Loc @ADVL
"<Ipmil>"
        "ipmil" N Sg Nom @SUBJ
"<sivdnidii>"
        "sivdnidit" V TV Ind Prt Sg3 @+FMAINV
"<almmi>"
        "albmi" N Sg Acc @OBJ
"<ja>"
        "ja" CC @CC-NP
"<eatnama>"
        "eana" N Sg Acc @OBJ
"<.>"
        "." CLB <<<
"<Eanan>"
        "eanan" N Sg Nom @SUBJ
"<lei>"
        "leat" V IV Ind Prt Sg3 @+FMAINV
"<ávdin>"
        "ávdin" A Sg Nom @SPRED
"<ja>"
        "ja" CC @CC-NP
"<guorus>"
        "guorus" A Sg Nom @SPRED
"<.>"
        "." CLB <<<
"<Seavdnjatvuohta>"
        "seavdnjatvuohta" N Sg Nom @SUBJ
"<govčča>"
        "gokčat" V TV Ind Prt Sg3 @+FMAINV
"<čiekŋalasa>"
        "čiekŋalas" N Sg Acc @OBJ
"<ja>"
        "ja" CC @CC-VP
"<Ipmila>"
        "ipmil" N Sg Gen @GN>
"<Vuoigŋa>"
        "Vuoigŋa" N Prop Plc Sg Nom @SUBJ
"<sattáhalai>"
        "sattáhallat" V IV Ind Prt Sg3 @+FMAINV
"<čáziid>"
        "čáhci" N Pl Gen @GP>
"<bajábealde>"
        "bajábealde" Po @ADVL
"<.>"
        "." CLB <<<
"<De>"
        "de" Adv @ADVL
"<Ipmil>"
        "ipmil" N Sg Nom @SUBJ
"<celkkii>"
        "cealkit" V TV Ind Prt Sg3 @+FMAINV
"<:>"
        ":" CLB
"<”>"
        "”" PUNCT RIGHT
"<Šaddos>"
        "šaddat" V IV Imprt Prs Sg3 @+FMAINV
"<čuovga>"
        "čuovga" N Sg Nom @SUBJ
"<!>"
        "excl" CLB <<<
"<”>"
        "”" PUNCT RIGHT
"<Ja>"
        "ja" CC @CC-VP
"<čuovga>"
        "čuovga" N Sg Nom @SUBJ
"<šattai>"
        "šaddat" V IV Ind Prt Sg3 @+FMAINV
"<.>"
        "." CLB <<<
```

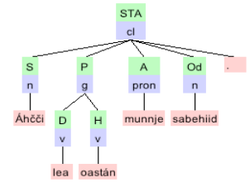
F r a m t i d s p e r s p e k t i v

Utvide til andre samiske språk

Vi kan bruke ein grammatisk analysator for eitt språk som utgangspunkt for tilsvarande analysatorar for nært beslekta språk. Det er seks samiske skriftspråk, vi har ein analysator for to av dei, er i ferd med å lage den tredje, og ser på dei tre siste. Delar av regeloppsettet kan brukast på nytt.

Pedagogisk programvare

Samiskundervisninga blir ofte organisert som desentralisert nettundervisning. Samiske verb blir bøygd i eintal, total og fleirtal i tre personar, ordet for «ikkje» er eit verb. Å lære dette mønsteret er ein sentral del av nybyrjaropplæringa, automaten vår gjer det mogleg å lage elevstyrte, interaktive spel som for det til.
Vi kan også la analyserte setningar gå inn i grammatikkspill, som her:



Informasjonssøk på samisk (Kva er vitsen med å lagre noko du veit du ikkje finn att?)

For å søke på ordformene *språk*, *språket*, *språkene*, *språka* kan vi skrive *språk**. For det tilsvarande samiske *giella*, er ca. 60 % av belegga i løpande tekst i såkalla svakt stadium (*giela*), mens ca. 10 % står i illativ (*gillii*). Berre 30 % står i sterkt stadium, dvs. at ved å søke etter *giella** går vi glipp av 70 % av dei relevante formene. Automaten vår kjenner att samtlege bøyingsformer av eit kvart samisk ord, og kan danne stammen i avanserte informasjonssøkeprogram.

586 Geavatlaččat mearkkaša dát ahte sámegiella ii leat doaibmi **giella** diehtojuohkinteknologiija oktavuođain .

680 Oslo universitehtas galgá sihkkarit ain leat oahpahus suoma-ugralaš **gielain** boahtteáiggis .

681 Dán áššis leai sáhka sámegiel fálaldaga heaittiheamis ii ge suoma-ugralaš **gielaid** heaittiheamis oppalaččat .

936 Dálá[lemma: giella / pos: N / syn: @OBJ / number: Pl / case: Acc]ohta sáhttá leat sihke áittan ja vejolašvuohtan sámi **giela** ja ku[illustreamis ja ovddideamis .

950 Kult[áilut lohkamuša girjjálašvuoda mas sámi mánát ja nuorat sihk[i seammás várjalivččii ja ovddidivččii sámi **giela** .

957 Sámediggi oaidná ahte kulturráđđi lea ruđaid prioriteren biergasiidda mat mánáidgárddiin galget geavahuvvot sámi **giela** ja kultuvrra gaskkusteapmái

S y n t a k t i s k e r e g l a r

```
"<Mii>"
        "M" Num Ill
        "mii" Pron Interr Sg Nom
        "mii" Pron Rel Sg Nom
        "mun" Pron Pers Pl1 Nom     ←  Fleirtalspronomen fordi
"<eat>"                                verbet står i fleirtal
        "ii" V IV Neg Ind Pl1

"<leat>"
        "leat" V IV Ind Prs Pl3
        "leat" V IV Ind Prs Sg2
        "leat" V IV Inf             ←  leat står etter nektignsverbet, det er altså
        "leat" V IV Ind Prs ConNeg  ←  ikkje Inf eller personbøygd form
        "leat" V IV Ind Prs Pl1
"<dan>"                                dan er ikkje demonstrativ, sidan det ikkje står foran
        "dat" Pron Pers Sg3 Acc
        "dat" Pron Dem Sg Acc       ←  adjektiv eller substantiv. Ikkje genitiv, sidan
        "dat" Pron Pers Sg3 Gen        det ikkje står etter talord eller preposisjon, og
        "dat" Pron Dem Sg Gen          ikkje foran substantiv, adjektiv eller postposisjon
"<muitalan>"
        "muitalit" V* TV Actio Der/eapmi N Sg Gen
        "muitalit" V TV Actio Acc
        "muitalit" V TV Actio Gen
        "muitalit" V TV Ind Prs Sg1
        "muitalit" V TV PrfPrc      ←  muitalan er partisipp, sidan det står
        "muitalit" V TV Actio Nom      som komplement til nektingsverb
"<.>"
        "." CLB
```

```
"<Mii>"
        "mun" Pron Pers Pl1 Nom @SUBJ
"<eat>"
        "ii" V IV Neg Ind Pl1 @+FAUXV
"<leat>"
        "leat" V IV Ind Prs ConNeg @-FAUXV
"<dan>"
        "dat" Pron Pers Sg3 Acc @OBJ
"<muitalan>"
        "muitalit" V TV PrfPrc @-FMAINV
"<.>"
        "." CLB <<<
```

Vi samarbeider med:
Tekstlaboratoriet, Universitetet i Oslo;
Oqaasileriffik, Nuuk; Visl, Syddansk universitet;
Kvensk institutt, Porsanger;
Institut för Språkvetenskap, Univ. i Helsingfors.