Disambiguering av homonymi i nord- og lulesamisk

Trond Trosterud og Linda Wiechetek
Det humanistiske fakultet
Universitetet i Tromsø

1 Innleiing

Knut Bergsland sitt råd til nye generasjonars uralistar var: Vi må stå på kvarandres skuldre, og ikkje på kvarandres tær. For vår del er skuldrene vi står på Pekka sitt samarbeid med Kimmo Koskenniemi om ein samisk analysator på 80- og særleg tidleg på 90-talet, som dannar basisen for arbeidet vårt med å disambiguere homonymi i samisk tekst.

I nordsamisk løpande tekst har kvar ordform i gjennomsnitt 2,6 moglege grammatiske analyser, og i lulesamisk har kvar ordform 2,0 moglege analyser.[1] Grunnen til at nordsamisk har meir homonymi enn lulesamisk er framfor alt at lulesamisk, fonologisk og morfologisk sett, er eit meir konservativt språk enn nordsamisk. Der nordsamisk final *-p, -t, -k* har falle saman i *-t*, og akkusativ og genitiv er identiske (bortsett frå i talord), har lulesamisk halde oppe desse distinksjonane. Den lulesamiske systematiske homonymien genitiv sg = nominativ pl er ikkje like omfattande som dei nordsamiske.

I begge språka er lokativ/inessiv pl. identisk med komitativ sg. Tabell 1 viser bøyingsmønsteret for det samiske substantivet *giella* "språk".

Tabell 1: Nordsamisk og lulesamisk kasusparadigme

Kasus	Sg	Pl	Kasus	Sg	Pl
Nom	giella	gielat	Nom	giella	giela
Akk	giela	gielaid	Akk	gielav	gielajt
Gen	giela	gielaid	Gen	giela	gielaj
Ill	gillii	gielaide	Ill	giellaj	gielajda
Loc	gielas	gielain	Ine	gielan	gielajn
			Ela	gielas	gielajs
Com	gielain	gielaiguin	Com	gielajn	gielaj
Ess	giellan		Ess	giellan	

I tillegg til homonymiane som er vist her er det også ei relativt lita gruppe ord, både i nord- og lulesamisk, som ikkje har stadieveksling, eller der stadievekslinga er synleg berre i tale, og ikkje i skrift. For desse orda blir homonymifelta større, t.d. blir nominativ = genitiv = akkusativ i nordsamisk for orda i figur 1, som har fått eit eige sett, figur 1 (nokre av desse orda kan også ha stadieveksling).

Når det gjeld verba (tabell 2), har nordsamisk ein systematisk homonymi mellom infinitiv og 1. person pluralis presens. For trestava verb representer den same forma også 3. person pluralis og 2. person singularis preteritum, alle er *muitalit*. For lulesamisk er dette tre ulike former: Infinitiven

[1]Vi vil takke Marit Julien for det arbeidet ho har gjort med utarbeidinga av regelsettet for nordsamisk. Vi vil også takke kollegene våre i Giellatekno- og Divvun-prosjekta for det kollektive arbeidet vi alle er ein del av: Lene Antonsen, Børre Gaup, Saara Huhmarniemi, Ilona Kivinen, Sjur Moshagen, Thomas Omma, Maaren Palismaa og Tomi Pieski.

```
LIST NOM-GEN-NOUN = "ILO-tráktáhta" "TV-kamera" "agens" "agitáhtor" "ahkit"
"akkumuláhtor" "aksiálláger" "aktuáhtor" "album" "autoritehta"
"journála" "lága" "pláhta" "plána" "stáhta" "ulbmil" "undulláhta" (...) ;
```

Figur 1: Sett av ord i nordsamisk som ikkje har stadieveksling

er *mujttalit*, 1. person pluralis har halde på opprinneleg *-p* og er *mujttalip*, mens dei to andre grammatiske orda har same form, *mujttali*.

Tabell 2: Nord- og lulesamiske verb i indikativ

Person	Nordsamisk *muitalit*		Lulesamisk *mujttalit*	
	Presens	Preteritum	Presens	Preteritum
Sg1	muitalan	muitalin	mujttalav	mujttaliv
Sg2	muitalat	muitalit	mujttala	mujttali
Sg3	muitala	muitalii	mujttal	mujttalij
Du1	muitaletne	muitaleimme	mujttalin	mujttalijma
Du2	muitaleahppi	muitaleidde	mujttalihppe	mujttalijda
Du3	muitaleaba	muitaleigga	mujttalibá	mujttalijga
Pl1	muitalit	muitaleimmet	mujttalip	mujttalijma
Pl2	muitalehpet	muitaleiddet	mujttalihpit	mujttalijda
Pl3	muitalit	muitaledje	mujttali	mujttalin

Når vi høyrer samisk tale er vi nesten aldri i tvil om kva form talaren har meint. Grunnen til det er at orda aldri blir ytra i isolasjon, men i ein kontekst som gjer at berre ei tolking er mogleg. I tillegg til dette har vi tilgang til både syntaktisk og semantisk informasjon. Likevel finst det enkelte tilfeller av ekte ambiguitet, spesielt når konteksten ikkje er tilstrekkelig. To eksempel med komitativ/lokativ homonymi: *1. oktiigullevašvuođa sámi servodagain* - a) "samhørighet/solidaritet med det samiske samfunnet" b) "solidaritet i de samiske samfunnene" *2. mo aššiin manná?* - a) "korleis går det med saka?" b) "korleis går det i sakene¿'

Vi vil gå inn på måtar å formalisere desse kontekstane på. Ein sentral konklusjon vil bli at formaliseringa kan bli reint grammatisk berre til eit visst punkt, og at vi deretter er avhengig av semantisk og leksikalsk kontekst for å kunne disambiguere rett. Desse resultata er relevante for lingvistiske modellar som hevdar å kunne modellere dei kognitive prosessane som trengst for å forstå språklege ytringar.

2 Homonymi i løpande tekst

2.1 Lulesamisk

Ei analyse av eit lulesamisk korpus på 136960 ord (det lulesamiske nytestamentet) gjev 259648 analyser, eller 1,9 analyser per ordform. Jf. tabell 3 for eit oversyn over fordelinga av homonymien.

Vanlegast er homonymien mellom eintal genitiv og fleirtal nominativ, som vi såg i tabell 1. Deretter kjem eit sett av verbformer, 2815 av dei 2867 tilfella er copula *le*. Det neste tilfellet har med namneformatering å gjere. Vi har kasuskongruens mellom fleirtalsformene, genitiv, komitativ, og, for visse stammeklasser, illativ. Interrogative og relative pronomen er identiske, dette er eit tilfelle av funksjonsdeling og ikkje eigentleg homonymi. Personlege pronomen skil ikkje mellom

genitiv og akkusativ. Blant verba er tredje person pluralis for visse stammeklasser identisk med første person dualis, for andre med preteritum andre person singularis.

Tabell 3: Dei sju vanlegaste homonymiklassene i lulesamisk løpande tekst

% hom	# hom	homonymiklasser
4,21%	6125	N Sg Gen = N Pl Nom
1,97%	2867	V Ind Prs Sg2 = V Ind Prs ConNeg = V Ind Prs Sg3 = V Ind Prs Pl3
1,20%	1746	N Pl Gen = N Pl Com
1,14%	1653	Pron Interr Sg Nom = Pron Rel Sg Nom
1,13%	1640	Pron Pers Sg3 Gen = Pron Pers Sg3 Acc
1,09%	1585	V TV Ind Prt Pl3 = V TV Ind Prs Du1
1,06%	1543	V TV Ind Prs Pl3 = V TV Ind Prt Sg2

2.2 Nordsamisk

Eit tilsvarande nordsamisk korpus (nytestamentet) på 165885 ordformer gjev 389524 analyser, eller 2,35 analyser per ordform, jf. tabell 4.

Den vanlegaste homonymien er den mellom akkusativ og genitiv, desse kasusa er alltid identiske. Homonymien ligg altså eigentleg i syntaksen, ikkje i kasusa, og vi må seie at det er akkusativobjekt og genitivmodifikator som er alternative analyser. Viss vi ignorerer fleirbruksfunksjonen til interrogative og relative pronomen, er dei første tre plassane på tabellen opptekne av akkusativ/genitiv-homonymi, meir enn ein tidel av homonymitilfella er av denne typen. Andre store grupper er lokativ *-s*, som er identisk med possessivsuffikset i genitiv og akkusativ, og infinitiv og presens første person fleirtal, som alltid er identiske.

Tabell 4: Dei ni vanlegaste homonymiklassene i nordsamisk løpande tekst

% hom	# hom	honomnymiklasser
5,27%	7670	N Sg Acc / N Sg Gen
2,27%	3308	N Sg Nom / N Sg Acc / N Sg Gen
1,82%	2645	Pron Interr Sg Nom / Pron Rel Sg Nom
1,78%	2583	N Pl Gen / N Pl Acc
1,73%	2516	Pron Pers Sg3 Nom / Pcle
1,56%	2275	N Sg Loc / N Sg Acc PxSg3 / N Sg Gen PxSg3
1,53%	2228	Pcle / CS
1,34%	1955	V TV Ind Prs Sg3 / V TV VGen / V TV Ind Prs ConNeg / V TV Imprt Prs Sg2 / V TV Imprt Prs ConNeg
1,26%	1840	V IV Inf / V IV Ind Prs Sg2 / V IV Ind Prs Pl1 / V IV Ind Prs ConNeg / V IV Ind Prs Pl3

3 Allment om disambiguering

Når vi skal disambiguere står vi ovafor i prinsippet to ulike slags tilfelle: Tilfella der dei homonyme formene opptrer i ulik grammatiske kontekst, og tilfelle der dei opptrer i lik grammatiske kontekst. Medlemmane i same homonymiklasse kan opptre både i like og i ulike grammatiske kontekstar. Genitiv og akkusativ er døme på dette: Som utvetydig objekt er det akkusativ, som utvetydig postposisjonskomplement er det genitiv (ulik kontekst), og som tidsadverbial kan det vere begge delar (lik kontekst). Dessutan kan det vere tilfelle der elles distinkte kontekstar er nøytralisert: I ein

streng *TV X Y Z Po*, som t.d. *Mun oainnán Biera biilla duohken.* "Eg ser Pers bils dørs.bak", kan vi ikkje vite om eg ser Per bak bildøra, eller Pers bil bak døra, dvs. vi veit ikkje kor akkusativen skal vere (på X eller Y), det kan til og med vere at setninga ikkje har akkusativ i det heile, at eg driv og ser meg rundt attom bildøra til Per. Uten pålitelig semantisk eller kontekstuell informasjon står vi ovafor eit tilfelle av ekte ambiguitet, der ikkje eingang eit menneske er i stand til å disambiguere.

3.1 Ulike innfallsvinklar til parsing

Det finst mange innfallsvinklar til disambiguering av homonymi i tekst. Dei to viktigaste er statistisk og grammatisk disambiguering. Statistisk basert disambiguering tar utgangspunkt i eit korpus av manuelt tagga tekst. Taggaren lærer mønster av dette korpuset, og går ut i frå at ny tekst oppfører seg på same måten. Ulempa med statistiske disambiguatorar er at dei ser ut til å ha eit tak på i underkant av 97 % korrekte resultat. Grammatisk baserte innfallsvinklar kjem i to versjonar, ovanfrå og ned, eller nedanfrå og opp. Ovanfrå-og-ned-taggarar (typisk: LFG-taggarar) prøver ut hypotesar om kva syntaktisk trestruktur setninga kan representere. Viss taggaren klarer det, er resultatet svært godt: ein syntaktisk frasestruktur med opplysningar om form, funksjon og hierarkisk struktur. Problemet med slike taggarar er at dei er både for gode og for dårlege: Same setning kan representerast av mange ulike hierarkiske analyser, og ei og same setning kan dermed få hundrevis, ja titusenvis av ulike analyseframlegg. På den andre sida er naturleg språk ofte ikkje i samsvar med syntaktiske reglar, talarar endrar meining midt i setninga, dei føyer inn lange digresjonar, eller setninga kan rett og slett bli for kompleks. I slike tilfelle bryt ovanfrå-og-ned-parsarar saman, og er ikkje i stand til å generere ein S-node i det heile. Sjølv gode parsarar av denne typen klarer sjeldan å analysere meir enn 60 % av setningane i løpande tekst.

Den innfallsvinkelen vi arbeider etter er ein grammatisk basert nedanfrå-og-opp-parsar og dermed ikkje bunde av den øvre grensa på 97 %. På den andre sida er han, på same måten som statistiske parsarar, ein nedanfrå-og-opp-parsar, og dermed robust: Også setningsfragment og komplekse setningar blir analysert. Det finst ulike typar nedanfrå-og-opp-parsarar, vår modell baserer seg på *føringsgrammatikk*, eller "constraint grammar", og ser grammatikken som eit sett *føringar*, eller "constraints", på kva kontekstar dei ulike morfologiske analysene kan opptre i. Føringsgrammatikken går attende til Karlsson (1990), Karlsson et al. (1995), og har vorte vidareutvikla av Tapanainen (1996) (cg2), Bick (2000) (vislcg). I arbeidet med samisk bruker vi Eckhard Bick sin versjon *vislcg*, `http://sourceforge.net/projects/vislcg/`.

3.2 Formalisme

Regelformatet til føringsgrammatikken er relativt enkelt, reglane er skrive som ein restriksjon, eller ei føring, på eit mål i ein gjeve kontekst, som vist i eksempel 1.

(1) OPERASJON mål IF kontekst ;

Reglane virkar på ei og ei ordform, som er definert som posisjon *0* av regelen. Ordet til venstre er *-1*, to ord til venstre er *-2*, og to eller færre ord til venstre er **-2-* Regelformatet inneheld operasjonane *ADD, MAP, REMOVE, SELECT*, dessutan kontekstidentifikatoren *IF* og vilkårsoperatorane *BARRIER, LINK*. Reglane med *ADD* og *MAP* har også ein eigen målidentifikator *TARGET*.

SELECT og *REMOVE* er dei sentrale operatorane, dei vel og fjerner lesingar, og er inverse av kvarandre: *SELECT A* inneber at alle andre lesingar enn A blir fjerna. *MAP* og *ADD* er operasjonar som legg til nye taggar, i vårt tilfelle legg dei til syntaktiske taggar, basert på plasseringa konstituentane har i setninga. *BARRIER* blir brukt i lag med *-operatoren. Når disambigueraren skannar setninga etter eit vilkår (t.d. **-1 B*, eller "sjå etter *B* til venstre"), i 2, kan vi legge inn ei mogleg barriere (**-1 B BARRIER C* "med mindre du finn *C* før du kjem så langt"). Vilkårsoperatoren *LINK* fører oss vidare frå eit mål: *(IF *-1 (N Nom) LINK *-1 CC LINK -1 (N Nom)* ber oss om å leite etter eit substantiv til venstre, med ein konjunksjon ein stad til venstre for seg, som på si side att har eit nytt sybstantiv umiddelbart til venstre for seg att.

Tabell 5: Regelformatet i føringsgrammatikken

ADD	legg til ny tagg
MAP	legg til tagg
REMOVE	fjern lesing
SELECT	vel lesing
IF	introduserer kontekstavgrensande operatorar
0	posisjonen til målet for operatorane
-1	ein posisjon til venstre
$*2$	to eller fleire posisjonar til høgre
BARRIER	stoppar skanninga
LINK	fører skanninga vidare til eit nytt vilkår

(2) (a) SELECT A IF *-1 B ;
(b) SELECT A IF *-1 B BARRIER C ;
(c) SELECT V-PL IF *-1 (N Nom) LINK *-1 CC LINK -1 (N Nom)) ;

Eit døme på ein regel er 3 c, ein regel som vel lesinga (dvs. analysen) *Inf* viss vi eitt ord til venstre finn eit medlem av settet *VFIN*, som tidlegare er definert som settet av *V-MOOD* minus alle analyser som inneheld negasjonsforma *ConNeg*.

(3) (a) SET V-MOOD = Ind | Pot | Imprt | ImprtII | Cond ;
(b) SET VFIN = V-MOOD - ConNeg ;
(c) SELECT Inf IF (-1 VFIN);

3.3 Hierarkisk struktur i ein lineær modell

Eit problem med disambigueringa er å klare å estimere frasestrukturen i setninga, når formalismen berre ser ein streng. Vi kan t.d. ikkje seie *substantiv til venstre for konjunksjon til venstre for NP*, når analyseprogrammet berre ser ein streng, og ikkje ein frase (NP).

For å få til å operere med hierarkiske strukturar i ein flat streng generaliserer vi over delar av strengen. For å simulere NP-ar definerer vi t.d. eit komplementært sett $NPNH$, der NPNH står for "NOT-PRE-NP-HEAD", ved først å sjå på kva som kan stå prenominalt i NP, og så definere komplementet til dette settet, som ($WORD - PRE - NP - HEAD$, der $WORD$ er definert som settet av alle ord).

(4) SET PRE-NP-HEAD = (Prop Attr) | (Prop @PROP>) | A | (Pron Pers Gen) | (N Gen) | Num | Cmpnd | CC | (Pron Dem) | (Pron Refl Gen) | (Pron Indef) | (PrfPrc @AN>) | (PrfPrc @PrcN>) | PrsPrc | (A Ord) ;
SET NPNH = WORD - PRE-NP-HEAD | ABBR ;

Det er meir å seie om NP-struktur enn dette, det finst relativsetninger og andre komplekse appositioner og modifikatorer som konkurrerer med adverbial og objekt, men med $NPNH$-definisjonen dekkar vi ein svært stor del av alle tilfella.

3.4 Setning eller avsnitt som grense

Disambigueringsprogrammet skanner strengen mellom to grensemerke, desse er definert av brukaren. Definisjonen *DELIMITERS = "<.>" "<!>" "<?>" "<...>" ;* viser grensene våre for disambiguering av setningar. I tillegg til det har vi ein separat disambigueringsmodul som har avsnitt

heller enn setning som domene, og som har eit separat regelsett. Grensemerkesettet er *DELIMITERS = "<¶>";*. Dette gjør det mulig å inkludere kontekst- og diskurs informasjon som kjem før sjølve setninga. I preprosesseringa av teksten legg vil til symbolet ¶ attom kvart avsnitt. Avsnittsdisambigueraren blir brukt til å disambiguere homonymi vi ikkje har vorte kvitt med setningsdisambigueraren, der det kan hjelpe å sjå ut over eiga setning, som td.d. i pro-drop-tilfelle, der vi kan identifisere personen til det finitte verbet via ein passande antesedent i forrige setning. I prinsippet er det mogleg å ha eit både større og mindre vindauge enn dette, men vi opererer altså med setning og avsnitt.

4 Disambiguere grammatisk skilde kontekstar

Vi ser no på disambiguering der dei homonyme formene opptrer i kontekstar som det er mogleg å skilje frå kvarandre på grammatisk grunnlag.

4.1 Verbperson

I nordsamisk er første person pluralis alltid identisk med infinitiv, som vi såg i tabell 2. Ordforma *vuolgit* kan bety både "dra" og "eg dreg". I eksempel 5 ser vi ei enkel setning, *Mii fertet vuolgit* "Vi må dra", der begge verbformene kan vere både infinitiv og første person pluralis, men der disambigueraren er i stand til å skilje mellom dei.

(5) *Mii fertet vuolgit.*
Vi må dra.

Den morfologiske analysatoren gjev oss alle dei moglege analysene i figur 2.

```
"<Mii>"
        "M" Num Ill
        "mii" Pron Interr Sg Nom
        "mii" Pron Rel Sg Nom
        "mun" Pron Pers Pl1 Nom
"<fertet>"
        "fertet" V IV Ind Prs Sg2
        "fertet" V IV Ind Prs Pl1
        "fertet" V IV Inf
"<vuolgit>"
        "vuolgit" V* IV N Actor Pl Nom
        "vuolgit" V IV Inf
        "vuolgit" V IV Ind Prs Pl1
"<.>"
        "." CLB
```

Figur 2: Morfologisk analyse av *Mii fertet vuolgit*

Disambiguatoren fjerner dei irrelevante analysene, og returnerer analysen i figur 3.

Regelen som vel Pl1-lesinga av verbet *fertet* er vist i 6. Regelen ser etter eit mogleg pronomen til venstre (**-1 MII-PERS*), i leitinga etter dette pronomenet får vi ikkje møte på finitte verb eller skiljeteikn på vegen, og heller ikkje eit nytt verb i første person pluralis til venstre for pronomenet:

(6) SELECT (V Pl1) IF (*-1 MII-PERS BARRIER VFIN OR PUNCT LINK NOT *-1 V-PL1 BARRIER NOT-ADV-PCLE OR CLB);

Regelen som vel infinitiv for *vuolgit* (nummer 7) plukkar ut infinitivslesinga viss det står eit verb som tar infinitivskomplement til venstre for seg, (**-1 INF-VERB*).

```
"<Mii>"
        "mun" Pron Pers Pl1 Nom @SUBJ
"<fertet>"
        "fertet" V IV Ind Prs Pl1 @+FAUXV
"<vuolgit>"
        "vuolgit" V IV Inf @-FMAINV
"<.>"
        "." CLB <<<
```

Figur 3: Analyse av den nordsamiske setninga *Mii fertet vuolgit*

(7) SELECT Inf IF (*-1 INF-VERB)(NOT *1 Inf BARRIER VFIN OR CS OR CP);

Verbet *fertet* "måtte" er eit slikt verb, eitt blant mange, som vi kan sjå i settet $INF-VERB$ i figur 4.

```
LIST INF-VERB = "astat" "ádjánit" "áigut" "álgit" "ásahit" "bágget"
"bávččagit" "beassat" "berret" "bivdit" "bivvat" "bistit"
"boahtit" "bovdet" "čohkkát" "čohkkedit" "čohkkánit" "čuoččahit"
"čuoččastit" "čuorvut" "čurggodit" (...) ;
```

Figur 4: Settet INF-VERB

Subjektet kan sjølvsagt vere meir komplekst; i og med at vi analyserer autentiske setningar har vi ofte svært kompliserte konstruksjonar. Eksempel 8 viser eit litt meir komplisert subjekt, koordinert personleg pronomen og NP:

(8) *Mun ja Maria váhnemat leat boahtán.*
Eg og Marias foreldre er komne.

Setninga i 8 får analysen i figur 5.

```
"<Mun>"
        "mun" Pron Pers Sg1 Nom @SUBJ
"<ja>"
        "ja" CC @CC
"<Maria>"
        "Maria" N Prop Fem Sg Gen @GN>
"<váhnemat>"
        "váhnen" N Pl Nom @SUBJ
"<leat>"
        "leat" V IV Ind Prs Pl1 @+FAUXV
"<boahtán>"
        "boahtit" V IV PrfPrc @-FMAINV
"<.>"
        "." CLB <<<
```

Figur 5: Analyse av *Mun ja Maria váhnemat leat boahtán*

Regelen som gjev oss Pl1 for det finitte verbet er regel 9.

(9) SELECT (V Pl1) IF (*-1 Nom BARRIER NOT-ADV-PCLE LINK *-1 CC BARRIER NPNH LINK -1 (Pers Sg1 Nom) OR (Pers Du1 Nom) OR (Pers Pl1 Nom));

Regel 9 vel Pl1 viss det til venstre finst ein nominativ, utan at det er noko anna enn adverb eller partikkel imellom, slik at denne nominativen er lenka til ein konjunksjon foran NPen (derfor BARRIER NPNH), til venstre for konjunksjonen skal vi finne eit pronomen i nominativ.

4.2 Genitiv vs. akkusativ i nordsamisk

Genitiv og akkusativ opptrer ofte i skilde kontekstar. For at vi skal ha akkusativ må vi enten ha eit tidsadverbial eller eit transitivt verb. Vi har gått gjennom alle verba i det nordsamiske leksikonet vårt og merka dei for transitivitet. Det er sjølvsagt mogleg å bruke eit transitivt verb intransitivt, som t.d. den samiske omsetjinga av *Et du hos Per*, der *Per* ikkje skal ha akkusativ, trass i at det er komplementet til eit transitivt verb.

Komplementet til ein postposisjon skal vere i genitiv. Possessoren til eit substantiv skal også vere i genitiv, denne possessoren kan stå til både eit postposisjonskomplement og eit objekt. Dermed kan vi ha både $VGGGP$ (med intransitiv bruk av verbet, $VAGGP$, $VGAGP$ og (i tilfelle postposisjonen er eit adverbial) $VGGAP$.

Det hendar strengen av substantiv gjev vink om den rette løysinga, t.d. viss eit av substantiva er eit eigennamn er det sannsynlegvis possessor heller enn det er modifisert av eit fellesnamn. Jfr. 10.

(10) *Mun oainnán beatnaga Máreha uvssa duohken.*
Eg ser hund.GA Marit.GA dør.GA bak.
'Eg ser (ein hund bak Marits dør. / ?? hundens Marit bak døra.)'

Mun oainnán beatnaga nieidda uvssa duohken.
Eg ser hund.GA jente.GA dør.GA bak.
'Eg ser (ein hund bak jentas dør. / hundens jente bak døra.)'

Mun oainnán beatnaga gusa uvssa duohken.
Eg ser hund.GA ku.GA dør.GA bak.
'Eg ser (ein hund bak kuas dør. / hundens ku bak døra.)'

Dei første to setninga klarer disambigueraren vår å disambiguere, jf. figur 6.

```
"<Mun>"
        "mun" Pron Pers Sg1 Nom @SUBJ
"<oainnán>"
        "oaidnit" V TV Ind Prs Sg1 @+FMAINV
"<beatnaga>"
        "beana" N Sg Acc @OBJ
"<Máreha>"
        "Máret" N Prop Fem Sg Gen @GN>
"<uvssa>"
        "uksa" N Sg Gen @GP>
"<duohken>"
        "duohken" Po @ADVL
"<.>"
        "." CLB <<<
```

Figur 6: Analyse av den nordsamiske setninga *Mun oainnán beatnaga Máreha uvssa duohken.*

Det som gjev oss akkusativ for *beatnaga* i dei to første setningane er nett det faktum at neste ord refererer til eit menneske, enten eigennamnet *Máret* eller substantivet *nieida*, jf. 11, der settet $HUMAN-INDIVIDUAL$ inneheld både fornamn og etternamn, og i tillegg namn for menneske (*bárdni, mánná, nieida, nisson, skihpár....*

(11) REMOVE Gen IF (0 (N Acc) LINK NOT 0 HUMAN-INDIVIDUAL OR PROFESSION OR OFFICE OR ROLE OR ("bearaš") OR NATION)(*1 HUMAN-INDIVIDUAL BARRIER NOT-ADJ LINK NOT 0 ("eadni") OR ("áhčči"));

I det siste tilfellet har vi ikkje nokon hjelp frå den semantiske statusen til substantiva, og strengen blir ikkje disambiguert, som vist i figur 7.

```
"<Mun>"
        "mun" Pron Pers Sg1 Nom S:2479 @SUBJ
"<oainnán>"
        "oaidnit" V TV Ind Prs Sg1 S:2336 @+FMAINV
"<beatnaga>"
        "beana" N Sg Acc @OBJ
        "beana" N Sg Gen @GN>
"<gusa>"
        "gussa" N Sg Acc @OBJ
        "gussa" N Sg Gen @GN>
"<uvssa>"
        "uksa" N Sg Gen @GP>
"<duohken>"
        "duohken" Po @ADVL
"<.>"
        "." CLB <<<
```

Figur 7: Analyse av den nordsamiske setninga *Mun oainnán beatnaga gusa uvssa duohken.*

Sjølv om det dreier seg om to grammatiske kasus med ulik posisjon i setningsstrukturen, har vi altså ofte tilfelle der vi ikkje finn rett svar. 10 er sjølvsagt konstruert, men slike konstruksjonar er ikkje uvanlege. Eit autentisk døme er 12.

(12) *SDD áigu deattuhit sámi álbmoga dárbbuid guovlluid*
SHD vil vektlegge same.Gen folk.Gen behov.Gen områda.Gen
dearvvašvuođafitnodagaid eaiggátstivremis.
helseforetaka.Gen eigarstyring.Loc.
'SHD vil legge vekt på den samiske befolknings behov i eierstyringen av de regionale helseforetakene.'

Her klarer heller ikkje programmet å gje noka analyse, vi får to analyser ståande att heile vegen.

4.3 Genitiv singularis vs. nominativ pluralis i lulesamisk

I lulesamisk skil akkusativ og genitiv seg frå kvarandre (bortsett frå i dei personlege pronomena). I staden er genitiv singularis identisk med nominativ pluralis. Dette er ein homonymi som er langt enklare å disambiguere, i og med at det ikkje berre er kasusskilnad, men også numerusskilnad, og i ot med at både genitiv og nominativ er grammatiske kasus som står i valensbunde tilhøve til kjerna i den frasen dei opptrer i. I eksempel 13 har vi ei enkel setning med eit subjekt og eit finitt verb. Både subjektet og verbet er tvetydige (verbet kan vere presens tredje person pluralis eller preteritum andre person singularis, og substantivet altså genitiv singularis eller nominativ pluralis).

```
"<SDD>"
        "SDD" N ACR Sg Gen @GN>
"<áigu>"
        "áigut" V TV Ind Prs Sg3 @+FAUXV
"<deattuhit>"
        "deattuhit" V TV Inf @-FMAINV
"<sámi>"
        "sápmi" N Sg Gen @GN>
"<álbmoga>"
        "álbmot" N Sg Acc @OBJ
        "álbmot" N Sg Gen @GN>
"<dárbbuid>"
        "dárbu" N Pl Acc @OBJ
        "dárbu" N Pl Gen @GN>
"<guovlluid>"
        "guovlu" N Pl Acc @OBJ
        "guovlu" N Pl Gen @GN>
"<dearvvašvuođafitnodagaid>"
        "dearvvašvuođa#fitnodat" N Pl Gen @GN>
"<eaiggátstivremis>"
        "eaiggát#stivret" V* TV Actio N Sg Loc @ADVL
"<.>"
        "." CLB <<<
```

Figur 8: Analyse av den nordsamiske setninga *SDD áigu deattuhit sámi álbmoga dárbbuid guovlluid dearvvašvuođafitnodagaid eaiggátstivremis.*

(13) *Ádjá* *mujttali*
Bestefar.GenPl fortel.Pl3
'Bestefedrene fortel'

Dei to orda i den lulesamisk setninga 13 har altså begge to moglege analyser, som vist i figur 9:

```
"<Ádjá>"
         "áddjá" N Pl Nom
         "áddjá" N Sg Gen
"<mujttali>"
         "mujttalit" V TV Ind Prs Pl3
         "mujttalit" V TV Ind Prt Sg2
"<.>"
         "." CLB
```

Figur 9: Moglege morfologiske analyser i den lulesamiske setninga *Ádjá mujttali*

Setninga har likevel berre ein mogleg analyse, og den lulesamiske disambigueraren vår finn denne analysen, jf. figur 10.

Reglane står i 14, den første seier at vi skal fjerne genitivlesinga viss nominativ pluralis er eit alternativ, viss det er eit verb i tredje person pluralis til høgre, og det ikkje er førstepersonspronomen som kan gje oss andre analyser til venstre. Regelen for Pl3 leitar etter sikre kandidatar

```
"<Ádjá>"
        "áddjá" N Pl Nom
"<mujttali>"
        "mujttalit" V TV Ind Prs Pl3
"<.>"
```

Figur 10: Analyse av den lulesamiske setninga *Ádjá mujttali*

for nominativ pluralis til venstre (C-en i $*-1C$ står for *careful mode*, dvs. vi er forsiktig og krev at substantivet allereie skal vere disambiguert, og det er det, reglane i 14 er **ordna**, og står i eit feeding-bleeding-tilhøve til kvarandre: Med motsett rekkjefølgje ville vi ikkje ha fått tilslag.

(14) (a) REMOVE (N Sg Gen) IF (*-1 BOS OR CS BARRIER (Pron Pers Du1))(0 (N Pl Nom))(1 (V Ind Pl3) LINK NOT 0 Imprt OR ImprtII);

(b) SELECT Pl3 IF (*-1C (N Pl Nom) BARRIER (Pron Pers Sg2 Nom) OR (Pron Pers Sg3 Nom) OR (N Sg Nom));

5 Disambiguere homonymi som opptrer i same grammatiske kontekst

Typiske døme på homonymi i same kontekst er adverbial, som er laust knytt til setninga (adjunkt), slik at fleire analyser er moglege. I nordsamisk er det komitativ singularis vs. lokativ pluralis som er det beste dømet på homonymi i same grammatiske kontekst.

5.1 Lokativ pluralis vs. komitativ singularis

Lokativ pluralis og komitativ singularis er alltid identiske i nordsamisk, uavhengig av variasjon i stammeklasse og bøyingsklasse. Begge kasusa er adverbiale kasus, og i dei fleste tilfella kan begge to i utgangspunktet opptre i same setning. Sjølv om den eine kasusforma kan vere valensbunde (oblik adverbial), kan den andre i teorien også opptre enten i tillegg til det valensbunde adverbialet eller aleine i absolutt bruk av predikatet. Jf. 15:

(15) *Dat ságastallet Sámedikkiin*
Dei diskuterer med Sametinget (valensbunde komitativ)
Dei diskuterer i Sametinga (absolutt bruk)

Det kan vere vanskeleg å avgjere om noko skal gjerast *med Sametinget* (komitativ singularis) eller *i Sametingene* (lokativ pluralis).

Det er nødvendig å vite litt om bruken av lokativ og komitativ i nordsamisk. Mens lokativen er ein svært fleirtydig kasus med mange bruksmolegheiter, er komitativen heller trang i forhold til bruken. Med bruk meiner vi her semantiske roller. Vi tar utgangspunkt i en forenkla versjon av rollesystemet i Sammallahti (2002),Sammallahti (2005) .

Lokativ har prototypisk ein tilstads- og fråstadsfunksjon, i tilegg står eigaren i ein habitiv-konstruksjon i lokativ, og lokativen brukast i ein del-av-relasjon. Lokativen kan få følgjande semantiske roller: MUHTTAŠUVVI (endrar), FÁDDÁ (tema), EAIGGÁT (eigar), GÁLDU (kjelde), BÁIKI (stad), HÁLDDAŠEADDJI (possessor), LUOBAHEADDJI (donator). Komitativen derimot er ein prototypisk "verktykasus. "Verktyetkan vere både menneskeleg (sosiativ) eller ikkje-menneskelig (instrumental). Den instrumentale tydinga tilsvarer rolla GASKAOAPMI (instrument), den sosiative tydinga tilsvarer vanlegvis VÁSIHEADDJI (opplevar). Nokre predikat, som

kan vere både verb og (mest deverbale) substantiv, skapar ein potensiell lokativkontekst, mens andre skapar ein komitativkontekst i høve til medlemer til eit spesifikt sett.

Vi har fleire strategiar for å angripe dette problemet. Den opplagte måten er å bruke subkategoriseringskriteria: Det finst rektio-verb som krev lokativ eller komitativ. Det følgjande settet er del av ei liste av verb som tar lokativ som komplement. Vanskelegare blir det med einingar som ikkje bare modifiserer predikatet, men heile setninga.

```
LIST LOC-VERB = "ávkkástallat" "ballat" "beassat" "beroštit" "biehttalit"
  "bihtit" "boahtit" "ceavzit" "čuoččut" "čuovvut" (...);

LIST COM-VERB = "árvalit" "árvvohuššat" "ávkašuvvat" "bargat" "bártašuvvat"
  "buohtastahttit" "deaivvadit" "háladit" "hilbošit" (...) ;
```

Figur 11: Sett av lokativ- og komitativverb

Regel 16 vel komitativ (og ekskluderer dermed lokativ) innafor ei heil- eller leddsetning viss det ikkje er ein potensiell habitiv konstruksjon i vegen (som da kunne kreve lokativ).

(16) SELECT Com IF (0 Sg)(*1 COM-VERB BARRIER CS OR CP LINK 0 VERB) (NOT *1 COPULAS BARRIER VERB LINK *1 COM-VERB BARRIER NOT-ADV-PCLE LINK 0 Inf);

Eit slikt døme er 17, der komitativen *eanadoaluin* blir disambiguert av regelen 16, jf. resultatet i figur 12.

(17) *... go sápmelaččat duođas álge eanadoaluin bargat.*
då samane på alvor byrja jordbruk.Com å arbeide
'Då samane på alvor byrja å arbeide i jordbruket'

```
"<...>"
        "..." CLB <<<
"<go>"
        "go" CS @CS
"<sápmelaččat>"
        "sápmelaš" N Pl Nom @SUBJ
"<duođas>"
        "duođas" Adv @ADVL
"<álge>"
        "álgit" V IV Ind Prt Pl3 @+FMAINV
"<eanadoaluin>"
        "eana#doallu" N Sg Com @ADVL
"<bargat>"
        "bargat" V TV Inf @-FMAINV
"<.>"
        "." CLB <<<
```

Figur 12: Analyse av *...go sápmelaččat duođas álge eanadoaluin bargat*

Disambiguering av komitativ og lokativ blir vanskelegare når kompleksiteten til setningane aukar. For lange tekstar omsett frå norsk som handlar om juridiske spørsmål, kan det vere vanskeleg å formalisere skilje mellom lokativ og komitativ. Problematikken kan ligge i omsetjinga: Norsk "med"-setning blir omsett med komitativ til samisk sjølv om det kanskje ikkje er den intuitive løysinga for samisk i utgangspunktet.

5.2 Singularis vs. pluralis

I visse tilfelle går det an å disambiguere mellom lokativ og komitativ med å utnytte at den eine står i singularis og den andre i pluralis. Det finst også ord som ut i frå semantikken ikkje har ei pluralisform viss den ikkje blir framheva eksplisitt med hjelp av ord som *máŋgalágan*, eit numeral eller eit ordenstal. Eigennamn er også substantiv som vanlegvis ikkje finst i pl. Andre singularisord kan vere ord som allereie er spesifisert som unik. Namn på språk står t.d. i eintal, og får dermed komitativ. Andre slike singularisord kan vere abstrakte ord som *ipmárdus* og *gelbbolašvuohta*.

Dei andre tilfella der det er mogleg å disambiguere singularis og pluralis går på konteksten. Nokre gjenstandar, objekt, abstrakte ting er unike i visse samanhengar, men ikkje i andre. Embetet "utanriksminister" finst det berre eitt av i Norge, men fleire i Europa. Tilsvarande gjeld for "fylkesting" innafor kvart einskild fylke. Jf. setninga i 18, regelen er vist i 19 og den resulterande analysen i figur 13

(18) Son deattuha sakka gulahallama sihke Sámedikkiin ja Finnmárkku fylkkadikkiin áššiˇ gieđahaladettiin.

(19) SELECT Sg IF (-1 FYLKA LINK 0 (@GN>))(0 ("fylkka#diggi" Com) OR ("fylkka#gielda" Com) OR ("fylkka#mánni" Com) OR ("fylkkas#gielda" Com) OR ("fylkkarehket#dárkun" Com))(NOT -3 FYLKA LINK -2 CC);

```
"<Son>"
        "son" Pron Pers Sg3 Nom @SUBJ
"<deattuha>"
        "deattuhit" V TV Ind Prs Sg3 @+FMAINV
"<sakka>"
        "sakka" Adv @ADVL
"<gulahallama>"
        "gulahallat" V TV Actio Acc @OBJ
        "gullat" V TV Der1 Der2 Der/halla Actio Acc @OBJ
"<sihke>"
        "sihke" Adv @ADVL
"<Sámedikkiin>"
        "Sámediggi" N Prop Org Sg Com @ADVL
"<ja>"
        "ja" CC @CC-NP
"<Finnmárkku>"
        "Finnmárku" N Prop Plc Sg Gen @GN>
"<fylkkadikkiin>"
        "fylkka#diggi" N Sg Com @ADVL
"<áššiˇ>"
        "áššiˇ" N Sg Gen @-FSUBJ
"<gieđahaladettiin>"
        "gieđahallat" V TV Ger @ADVL
"<.>"
        "." CLB <<<
```

Figur 13: Analyse av *Son deattuha sakka gulahallama sihke Sámedikkiin ja Finnmárkku fylkkadikkiin áššiˇ gieđahaladettiin.*

Visse ord, som abstrakte ord, opptrer sjelden i fleirtal. Vi testar ut ulike måtar å utnytte dette på, m.a. med settet i 14.

Ein av dei reglane som viser til dette settet er 20:

```
LIST SG-WORD = ("addit" Der/upmi) "almmolašvuohta" "anistupmi" "álgu" "áŋgirvuohta"
"ávvu" "ballu" "borakeahttáivuohta" (...) ;
```

Figur 14: Sett for ord som skal vere i singularis

(20) SELECT Sg IF (0 SG-WORD)(NOT *-1 Num OR Ord BARRIER NPNH)(*-1 BOS LINK NOT *1 (“goappeš”) OR (“goappašat”) OR (“earálágan”) OR (“máŋgalágan”));

Setninga, eller snarare tittelen, i 21, kan sjå ut til å ha ekte ambiguitet i forhold til komitativ og lokativ:

(21) *Buoret kvalitehtta sámi giella- ja kulturgelbbolašvuođain*
Betre kvalitet samisk språk- og kulturkompetanse.Com
’Betre kvaliteten med hjelp av / i samisk språk- pg kulturkompetanse(ar)’

Det som til slutt gjør at berre den andre versjonen (komitativ) er mogleg, er at kulturkompetanse bare finst i singular. Analysen blir dermed som i figur 15.

```
"<Buoret>"
        "buoredit" V TV Imprt Prs Sg2 @+FMAINV
"<kvalitehtta>"
        "kvalitehtta" N Sg Nom @SUBJ
"<sámi>"
        "sápmi" N Sg Gen @GN>
"<giella->"
        "giella" N SgNomCmp Cmpnd @CMPND
"<ja>"
        "ja" CC @CC
"<kulturgelbbolašvuođain>"
        "kultur#gelbbolašvuohta" N Sg Com @ADVL
```

Figur 15: Analyse av *Buoret kvalitehtta sámi giella- ja kulturgelbbolašvuođain*

6 Status quo

I arbeidet med å skilje mellom ulike adverbiale kasus tar vi mange ulike verkemiddel i bruk. Vi ser på dei leksikalske eigenskapane til verbet eller substantivet det fleirtydige ordet står til, og vi ser på leksikalske eigenskaper ved ordet sjølv. I tillegg utnyttar vi at formene skil seg frå kvarandre ikkje berre i kasus men også i numerus.

Sjølv om lulesamisk har mindre kasushomonymi enn nordsamisk, har det også homonymi mellom genitiv og komitativ fleirtal, som i *gielaj* i tabell 1. I dette tilfellet skil kasusa seg meir frå kvarandre enn for den nordsamisk komitativ-lokativ-homonymien, i og med at genitiv er eit grammatisk kasus og komitativ er eit adverbialt. På den andre sida har dei same numerus, noko som gjer det vanskelegare att.

Per oktober 2006 inneheld disambigueringsprogrammet vårt 321 reglar for tilordning av syntaktiske funksjonar, og 2 276 reglar for morfologisk disambiguering. Dei fjernar ikkje all homonymi, men etter at desse reglane har verka er biletet over homonymi eit heilt anna enn før disambigueringa. Til eit korpus av 2 330 998 ord gav disambigueraren [2] 2 466 842 analyser, eller 1 058 analyser per 1 000 ord. Dei vanlegaste morfologiske homonymiklassene som stod att var desse:

[2]Teksten var 3 årgangar av Min Áigi, 2003-2005, og disambigueraren som vart brukt var versjon 1.836

Tabell 6: Dei homonymiklassene i nordsamisk løpande tekst som står att etter disambiguering

% hom	# hom	homonymiklasser
0,15	5849	N Pl Loc @ADVL = N Sg Com @ADVL
0,10	3560	N Sg Gen @GN> = N Sg Acc @OBJ
0,07	2367	V IV Ind Prs Sg2 @+FAUXV = V IV Ind Prs Pl1 @+FAUXV = V IV Ind Prs Pl3 @+FAUXV
0,06	1466	N Pl Acc @OBJ = N Pl Gen @GN>
0,04	1437	CS @CS = Adv @ADVL
0,03	1017	V TV Ind Prt Sg2 @+FMAINV = V TV Ind Prs Pl3 @+FMAINV
0,03	756	V TV Ind Prt Sg2 @+FMAINV = V TV Ind Prs Pl3 @+FMAINV = V TV Ind Prs @+FMAINV
0,03	754	N Sg Gen @GN> = N Sg Nom @SUBJ
0,03	750	N Sg Gen PxSg3 @ADVL = N Sg Loc @ADVL

Samanlikna med biletet før disambiguering (tabell 4) er biletet no endra. Pl Loc vs. Sg Com, som var på 20. plass, og utanfor tabellen, er no størst. Ein annan homonymitype som relativt sett er vanlegare no er homonymien for verbperson, som har gått opp frå 9-10 plass til 3-6 plass. For dei andre er det mindre endringar: Genitiv vs. akkusativ er framleis dominerande, det same er CS (subjunksjon) vs. Adv (adverb).

Allment sett kan vi seie at dei homonymitilfella vi klarer best er dei som har ulik syntaktisk funksjon, slik som homonymi mellom ulike grammatiske kasus, og den mellom grammatiske og adverbiale kasus. Disambigueraren har mest problem med Pl Loc vs. Sg Com og med dei ulike verbposisjonane. Felles for desse er at dei homonyme formene har same syntaktiske funksjon, adverbial og finitt verb, og at det dermed er vanskeleg å disambiguere dei ut i frå posisjon.

7 Konklusjon

I denne artikkelen har vi vist at det er mogleg å disambiguere morfologisk homonymi med å sjå på den grammatiske og leksikalske konteksten orda står i. Vi har brukt føringsgrammatikk, ein grammatiske basert nedanfrå-og-opp-modell, heller enn alternativa, statistisk baserte nedanfrå-og-opp-modellar og grammatiske baserte ovanfrå-og-ned-modellar. Vi har vist at det er mogleg å få like gode resultat som for betre studerte språk. Vi har også vist at det er mogleg å oppno gode resultat for samisk disambiguering med ein modell som baserer seg på relativ posisjon i den lineæare strengen, trass i at samisk trandisjonelt har vorte rekna for eit språk med fri ordstilling. Våre resultat tyder på at i alle fall relevante delar av den samiske syntaksen har ei relativt fast ordstilling.

Sjølv om det er mogleg å formalisere konteksten grammatisk, er det mange tilfelle der grammatisk kontekst ikkje er nok til å disambiguere, som tilhøyrarar er vi avhengige av å forstå setninga for å velje rett analyse. I forsøka våre på å formulere denne forståinga på ein maskinlesbar måte, har vi konstruert sett som generaliserer semantikken til verb og substantiv. Med desse relativt enkle tiltaka har vi fått ein langt betre analyse. Vidare framskritt vil vere avhengig av ein semantisk analyse av heile leksikonet. I mange tilfelle vil også diskursen vere viktig for disambiguering. Den språklege diskursen kan vi ta med, ved å ta omsyn til større delar av teksten. Den utomspråklege diskursen kan vi til ein viss grad ta omsyn til, ved t.d. å gje programmet informasjon om tekstsjanger.

I den grad resultata våre har overføringsverdi til teoriar om språkleg kompetanse, ser vi at den grammatiske analysen tilhøyraren gjev av setninga han eller ho høyrer, er avhengig av den

leksikalske semantikken som er knytt til kvart einskild ord i setninga.

Referanser

Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Aarhus, 2000.

Fred Karlsson. Constraint grammar as a framework for parsing running text. In *13th International Conference on Computational Linguistics (COLING-90)*, pages 168–173, Helsinki, 1990.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Atro Anttila. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing. Mouton de Gruyter, Berlin, New York, 1995.

Pekka Sammallahti. Cealkkaráhkadus ja cealkkalahtut. *Sámi dieđalaš áigečála*, 1:29–58, 2002.

Pekka Sammallahti. *Láidehus sámegiela cealkkaoahpa dutkamii*. Davvi Girji, Kárášjohka, 2005.

Pasi Tapanainen. *The Constraint Grammar Parser CG-2*, volume 27 of *Publications of the Department of General Linguistics, University of Helsinki*. University of Helsinki, Helsinki, 1996.