

Samisk språkteknologi som pedagogisk programvare

Trond Trosterud
Universitetet i Tromsø
<http://giellatekno.uit.no/>



Samisk språkteknologi i dag

- Universitetet i Tromsø:
 - <http://giellatekno.uit.no>
 - analysere tekst, gjere teksten tilgjengeleg for forskning
 - lage pedagogisk programvare for samisk
 - Trond Trosterud, Marit Julien, Saara Huhmarniemi, Ilona Kivinen
- Sametinget og Universitetet i Tromsø
 - <http://www.divvun.no> (frå medio mai)
 - lage retteprogram for nord- og lulesamisk
 - Sjur Moshagen, Tomi Pieski, Børre Gaup, Thomas Omma, Maaren Palismaa

Oversikt over foredraget

1. Grammatisk analyse

a) morfologisk analyse av samisk

i. morfologisk analyse av einskildord

ii. morfofonologiske justeringsreglar

b) analyse av ord i kontekst (disambiguering)

2. Pedagogiske applikasjonar

3. Modellar for samarbeid om samisk språketeknologi

Oversikt over foredraget

1. Grammatisk analyse

a) morfologisk analyse av samisk

i. morfologisk analyse av einskildord

ii. morfofonologiske justeringsreglar

b) analyse av ord i kontekst (disambiguering)

2. Pedagogiske applikasjonar

3. Modellar for samarbeid om samisk språketeknologi

Grammatisk analyse

morfológisk
analyse

morfológisk
disambiguering

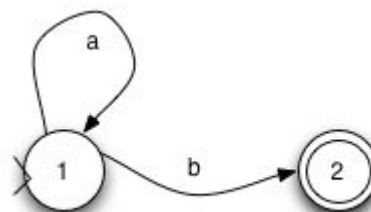
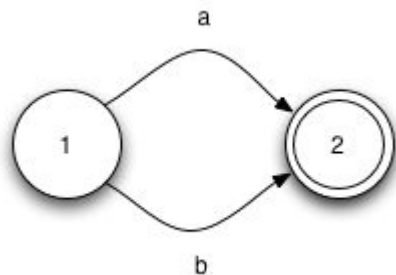
syntaktisk
analyse

Mun
mun Pron Pers Sg1 Nom
boran
borrat V Ind Prs Sg1
bora N Ess
bora N Sg Nom PxSg1
bora N Sg Gen PxSg1
bora N Sg Acc PxSg1
boran N Sg Gen
boran N Sg Acc
boran N Sg Nom
láibbi
láibi N Sg Acc
láibi N Sg Gen

Mun
mun Pron Pers Sg1 Nom
boran
borrat V Ind Prs Sg1
láibbi
láibi N Sg Acc

Mun
mun Pron Pers Sg1 Nom **SUBJ**
boran
borrat V Ind Prs Sg1 **VFIN**
láibbi
láibi N Sg Acc **OBJ**

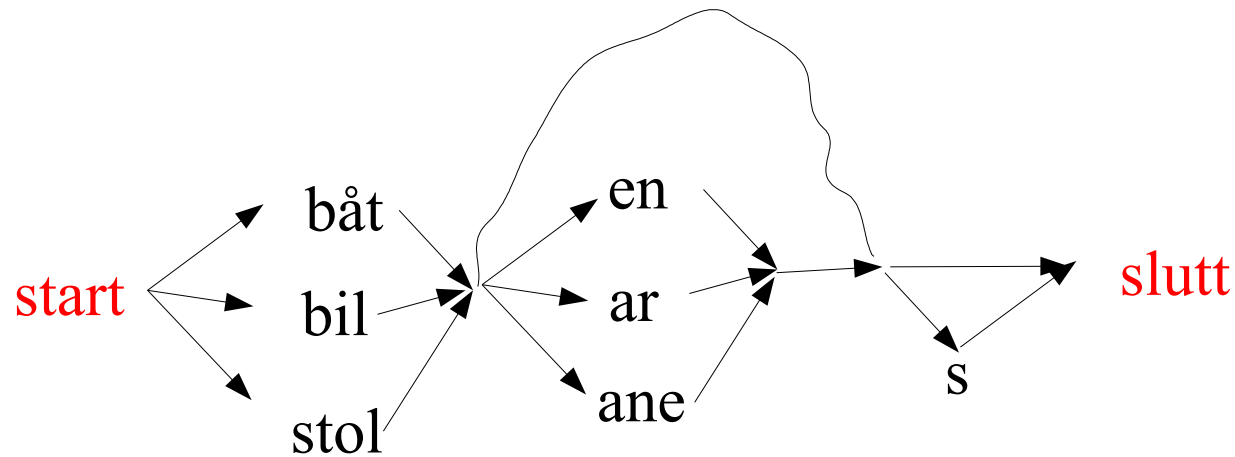
Automatar



Automatar aksepterer eller forkastar strengar
(dei er t.d. praktiske i retteprogram)

Automatar

- Morfologi som ein *endeleg tilstands-automat*
- Ein *automat* er ein sti frå starten til slutten av ordet



Morfologi som automatar

settet av alle ordformene i eit språk

$\langle = \rangle$

settet av dei stiane som fører frå start til slutt i
automaten

Det å representere morfologisk struktur som
endelege tilstandsautomatar er svært effektivt:

Komi:

$20000 \text{ substantiv} \times 16 \text{ Kasus} \times 2 \text{ Num} \times 6 \text{ Px} = 3,84 \text{ mill. former}$

$20000 \text{ substantiv} + 16 \text{ Kasus} + 2 \text{ Num} + 6 \text{ Px} = 20024 \text{ former}$

Transdusarar

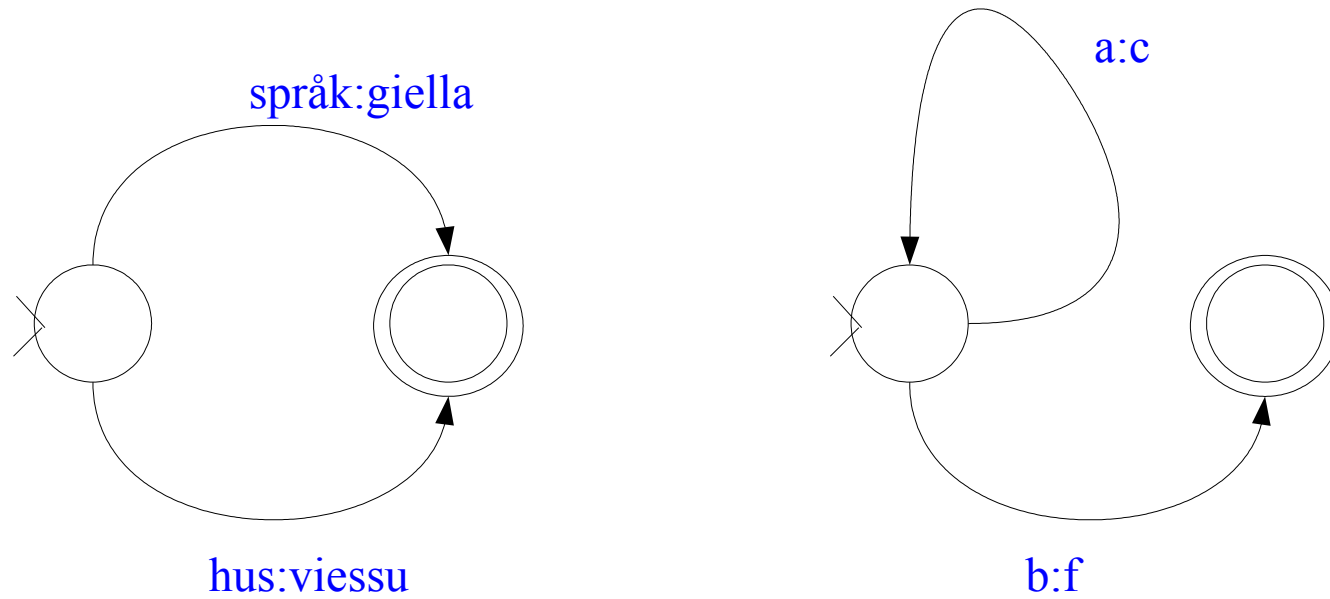
Automatar er maskiner som generer
symbolstrengar:

a, aaa, ababab, ...

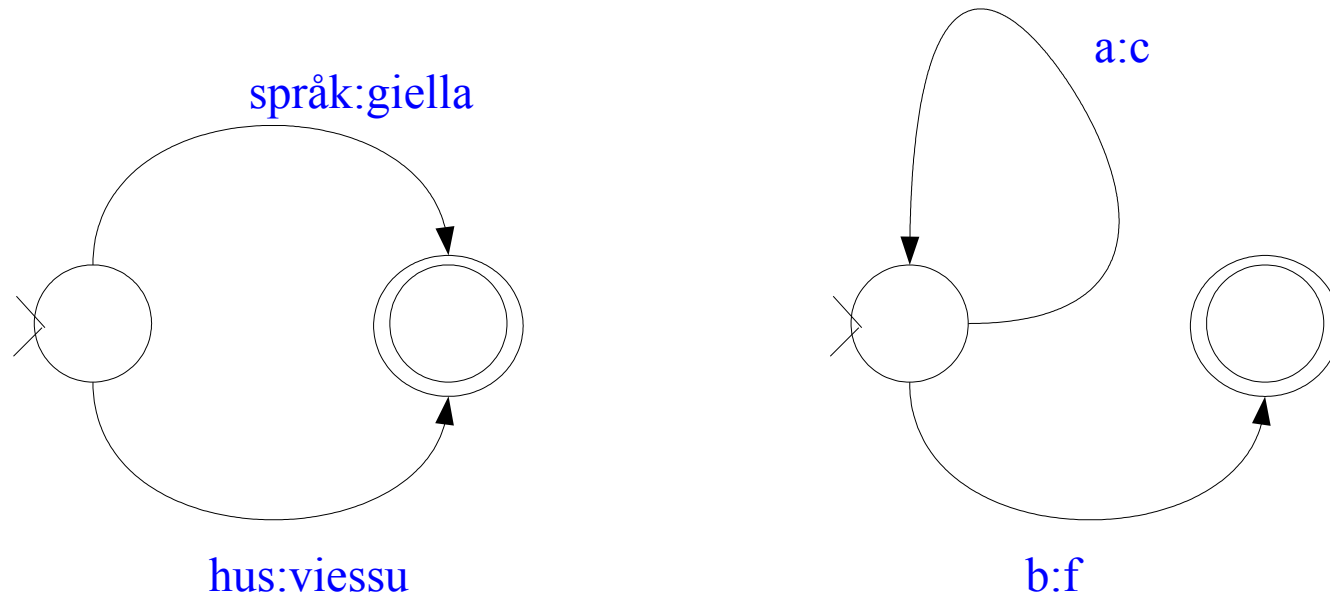
Transdusarar er automatar som genererer par av
symbol og symbolstrengar:

a:x, aaa:xxx, ababab:xyxyxy, ...

Transdusarar

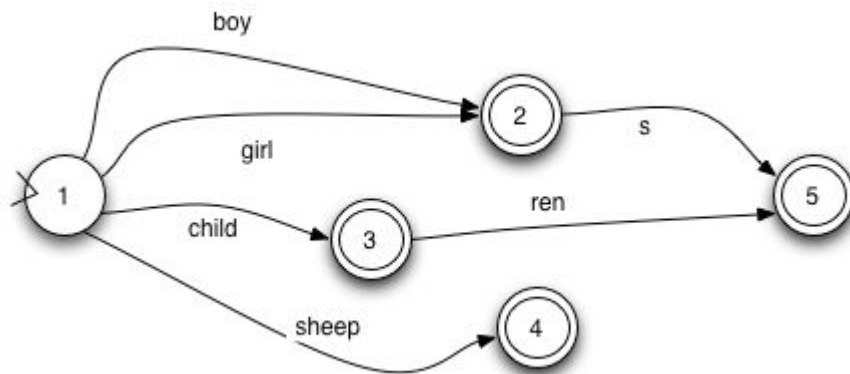


Transdusarar

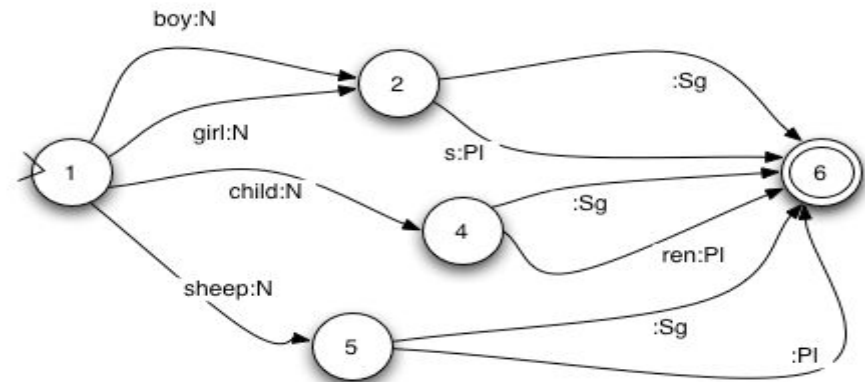


Transdusarar aksepterer (eller forkastar) inputstrengar strengar, og erstattar dei med andre strengar
(dei er t.d. praktiske til å lage ordbøker, eller grammatisk analyse)

Morfologiske automatar og transdusarar



Automat



Transdusar

Transdusarar kan koblast saman

Viss vi koblar transdusaren $a:b$ med transdusaren $b:c$ får vi ein ny transdusar $a:c$

$$\frac{\frac{a}{b}}{\frac{b}{c}} \Rightarrow \frac{a}{c}$$

Transdusar for ordbøying

morfo-
logi

leksem og grammatiske trekk

stamme med suffiks

muorra +N +Sg +Acc

muorra Q1 -v



morfo-
fono-
logi

stamme med suffiks

ordformer

muorra Q1 -v

muorav



Transdusar for ordbøying

morfo-
logi

leksem og grammatiske trekk

stamme med suffiks



morfo-
fono-
logi

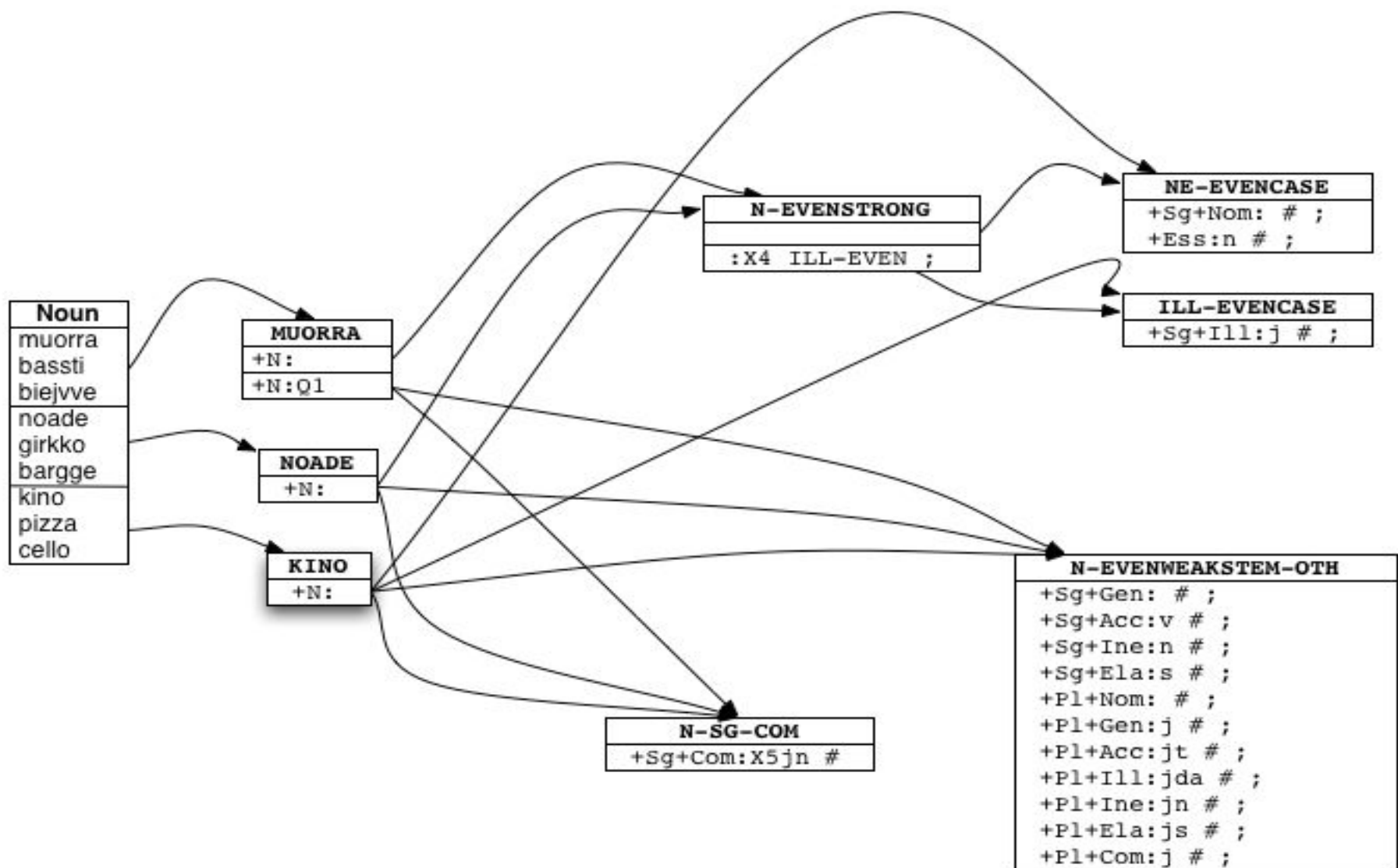
stamme med suffiks

ordformer

muorra +N +Sg +Acc

muorav

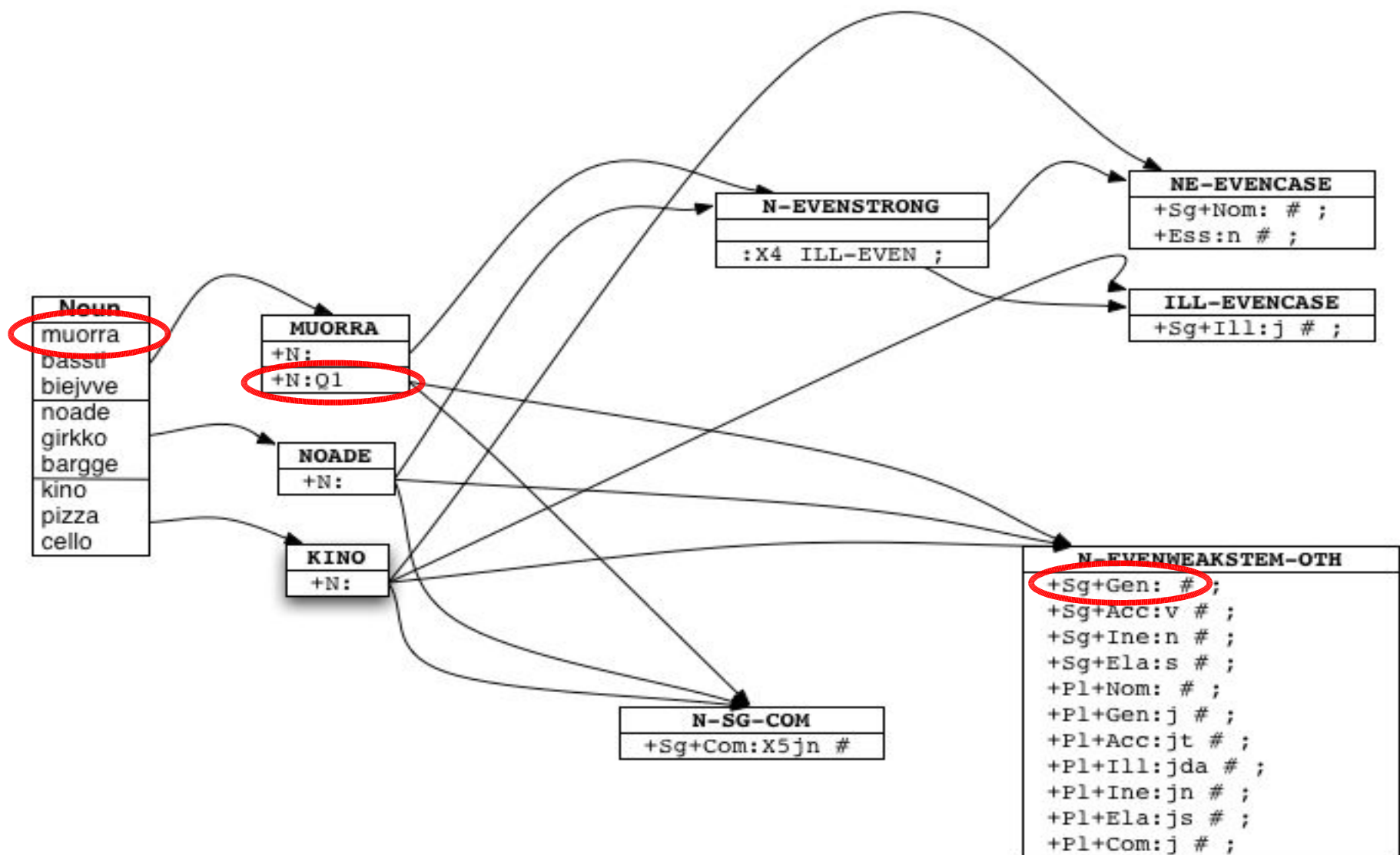
Morfologi: lulesamiske substantiv



Indre bøying med hjelp av kvasiaffiks

- Morfofonologiske prosessar blir kontrollert av arbitrære hjelpesymbol:
 - Q1 for stadieveksling
 - X4 for e:á og o:u i illativ sg.
 - X5 for e:i og o:u i komitativ sg.
- Desse hjelpesymbola blir sletta i løpet av prosessen, og er dermed usynlege for brukaren

Morfologi: lulesamiske substantiv

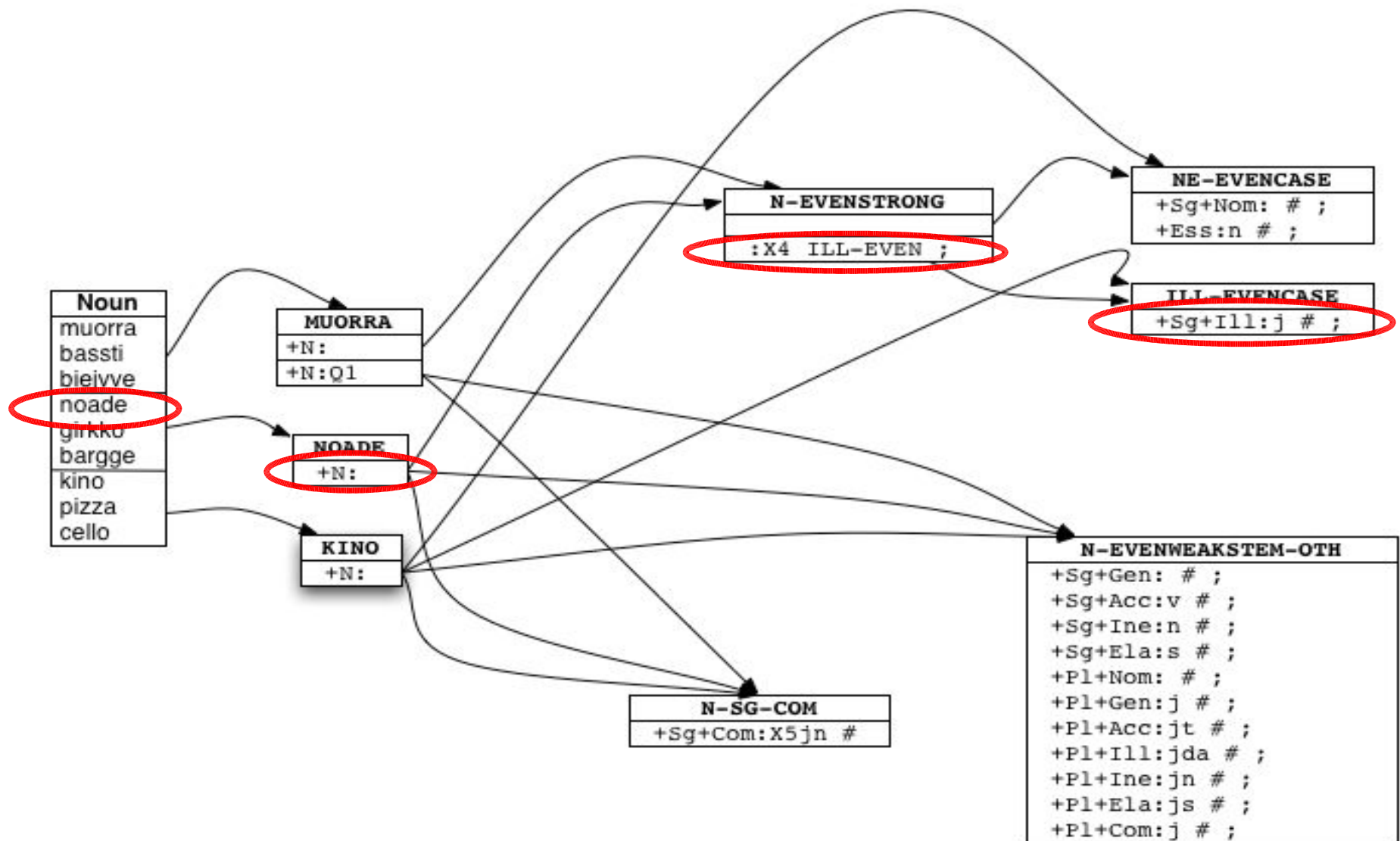


Stadieveksling foran Q1

r:0 <=> Vow: ([b|j|v]) _ r Vow: (StemCns:) Q1: ;

muorra+N+Sg+Gen	
muorraQ1	
<=>	
muorraQ1	
muo0ra	

Morfologi: lulesamiske substantiv



e:á foran X4 i illativ

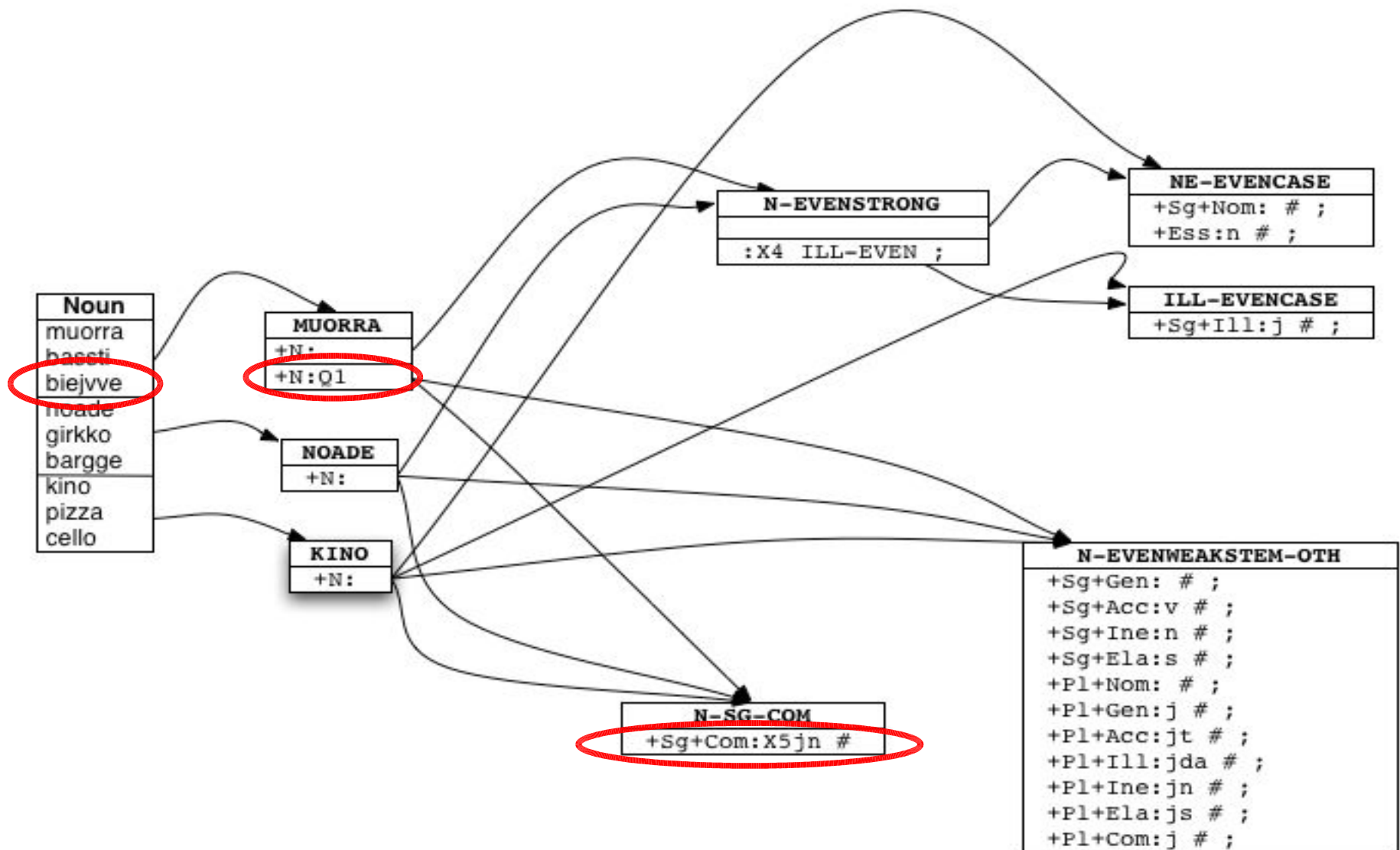
e:á <=> [Vow: - å] Cns* _ X4: ;

$$\frac{\text{noade+N+Ill}}{\text{noadeX4j}}$$

<=>

$$\frac{\text{noadeX4j}}{\text{noadå0j}}$$

Morfologi: lulesamiske substantiv



e:i foran X5 i komitativ sg.

e:i <=> Cns _ Q1:0 (X5:0) j [#: | Cns:] ;

v:0 <=> Vow: ([j|l|r]) _ v Vow: (StemCns:) Q1: ;

biejvve+N+Sg+Com

biejvveQ1X5jn

<=>

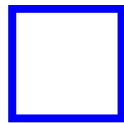
biejvveQ1X5jn

biej0vi0jn

Disambiguering

Ord i kontekst

"<Mun>"
"mun" Pron Pers Sg1 Nom
"<muitalan>"
"muitalit" V PrfPrc
"muitalit" V Act
 "muitalit" V Ind Prs Sg1



vel den lesinga som
inneheld Sg1

viss pronomenet *mun* står
ein stad til venstre

SELECT Sg1 IF *-1 ("mun" Pron Pers Sg1 Nom)
BARRIER Neg OR ("leat" V Ind Prs Sg1);

og det ikkje finst eit nektingsverb
eller verbforma *lean* før vi finn *mun*

"<Mun>"
"mun" Pron Pers Sg1 Nom
"<muitalan>" S:2196
"muitalit" V Ind Prs Sg1

"<Mun>"
"mun" Pron Pers Sg1 Nom
"<lean>" S:2200
"leat" V Ind Prs Sg1
"<muitalan>" S:2157
"muitalit" V PrfPrc

"<Mun>"
"mun" Pron Pers Sg1 Nom
"<in>"
"ii" V Neg Ind Sg1
"<lean>" S:2154
"leat" V PrfPrc
"<muitalan>" S:2154
"muitalit" V PrfPrc

Transitive setningar

"<Biret>"

"Biret" N Prop Sg Nom

"<liiko>"

"liikot" V VGen

"liikot" V Imprt Prs Sg2

"liikot" V Imprt Prs ConNeg

"liikot" V Ind Prs Sg3

"liikot" V Ind Prs ConNeg

"<skuvlii>"

"skuvlet" V Ind Prt Sg3

"skuvla" N Sg Ill

Transitive setningar

"<Biret>"

"Biret" N Prop Sg Nom

"<liiko>"

"liikot" V VGen

"liikot" V Imprt Prs Sg2

"liikot" V Imprt Prs ConNeg

"liikot" V Ind Prs Sg3

"liikot" V Ind Prs ConNeg

"<skuvlii>"

"skuvlet" V Ind Prt Sg3

"skuvla" N Sg Ill

REMOVE VGen IF (NOT *-1 V-IND-FIN
BARRIER CLB OR CS)
(NOT *1 V-IND-FIN
BARRIER CLB OR CS);

Transitive setningar

"<Biret>"

"Biret" N Prop Sg Nom

"<liiko>"

"liikot" V VGen

"liikot" V Imprt Prs Sg2

"liikot" V Imprt Prs ConNeg

"liikot" V Ind Prs Sg3

"liikot" V Ind Prs ConNeg

"<skuvlii>"

"skuvlet" V Ind Prt Sg3

"skuvla" N Sg Ill

REMOVE VGen IF (NOT *-1 V-IND-FIN
BARRIER CLB OR CS)
(NOT *1 V-IND-FIN
BARRIER CLB OR CS);

REMOVE Imprt IF (NOT -1 COMMA OR BOS or Pcle)
(NOT 0 (ConNeg)) ;

Transitive setningar

"<Biret>"
 "Biret" N Prop Sg Nom
"<liiko>"
 "liikot" V VGen
 "liikot" V Imprt Prs Sg2
 "liikot" V Imprt Prs ConNeg
 "liikot" V Ind Prs Sg3
 "liikot" V Ind Prs ConNeg
"<skuvlii>"
 "skuvlet" V Ind Prt Sg3
 "skuvla" N Sg Ill

REMOVE VGen IF (NOT *-1 V-IND-FIN
 BARRIER CLB OR CS)
 (NOT *1 V-IND-FIN
 BARRIER CLB OR CS);
REMOVE Imprt IF (NOT -1 COMMA OR BOS or Pcle)
 (NOT 0 (ConNeg)) ;
REMOVE ConNeg IF (NOT *-1 Neg);
REMOVE Sg3 IF (-1 VFIN)(0 (N Sg Ill));



"<Biret>"
 "Biret" N Prop Sg Nom
"<liiko>"
 "liikot" V Ind Prs Sg3
"<skuvllii>"
 "skuvlla" N Sg Ill

Oversikt over foredraget

1. Grammatisk analyse

a) morfologisk analyse av samisk

i. morfologisk analyse av einskildord

ii. morfofonologiske justeringsreglar

b) analyse av ord i kontekst (disambiguering)

2. Pedagogiske applikasjonar

3. Modellar for samarbeid om samisk språketeknologi

Fleire vegar å gå

- Vidareutvikle Tromsø-analysatoren/generatoren som ligg på nettet
- Samarbeide med andre miljø, og bruke vår programvare som input. Her: *visl*
 - <http://beta.visl.sdu.dk/>
 - <http://www.tekstlab.uio.no/grei/>

Tromsø

UiTs analysator/generator på nettet

- «Schimanski-modellen»: For å lese samisk lyrikk må ein ha tilgang til grunnformene. *bohtet* => **boahtit**
- Ein generator kan fortelje kva som er pret sg2 av verbet *boahtit*: **1. boahtit+V+Ind+Prt+Sg2 => bohtet**
- Mogleg vidareutvikling av analysatoren:
 - «Kjære datamaskin: Her har du dagens lekse, 15 nye verb. Kan du høyre meg i preteritumsbøyinga?»
 - «Ja, da, kom med dei. I tillegg skal du få ein grafisk statistikk som viser progresjonen din når du er ferdig...»

Bruk av språkressursane for utanforståande

- Analysere samisk via internett
 - Gå til <http://giellatekno.uit.no>
 - det er mogleg å lime inn tekst frå alle program (viss det er skrive i Unicode; Mac OS 9-brukarar må skrive **c1, d1, n1** for č, đ, ŋ)
 - lett å bruke
 - treng ikkje brukarløyve
 - Men:
 - analysatoren tar maks. 50 ord i slengen, for å unngå hackarar (kvantitativ analyse føreset brukarløyve)

Davisámegiela sátneanalysaprográmma

←

→

↺

+

http://giellatekno.uit.no/d-sme.html

Google

Osoitekirja ▾

G

AlltheWeb

koti

gt-i

gt-e

bibsys

ordbok

uit-haku

puh-luett

gule

Dál sáhtát analyseret sámegiel sániid

Sámegielbustávid čálát ná: Bustáva "á" don sáhtát čállit dego "á". Eará bustávid fertet čállit "c1, d1, n1, s1, t1, z1", sáni čállá don fertet ovdamearkka dihtii čállit ná: "c1állá". Prográmma addá sátnehámi morfologalaš analiissa. Čilgehus giellaoahpalaš oanádussii gávnnat [dás](#).

Jos válljet "Analysere", de prográmma addá buot analiissaid, muhto jos válljet "Disambiguere", de prográmma geahččala addit dušše analiissa, mii heive dan cealkagii.

Here you may analyse Northern Sámi words
The Sámi special letters are written like this: "á" is written like "a", and the other six letters with basic letter + the digit 1, like this: "c1, d1, n1, s1, t1, z1". Write the words in the window, and press the "Analyse" or "Reset" (= "Sihko") button. Grammatical tags are explained [here](#).

If you choose "Analysere", the program will give all possible analyses, whereas if you choose "Disambiguere", the program will try to give only the analyses that are appropriate for the given sentence.

Čále sátnehámi:

Mun háliidan c1állit sámegiela

☐ Analysere

☒ Disambiguere

Analysere

Sihko

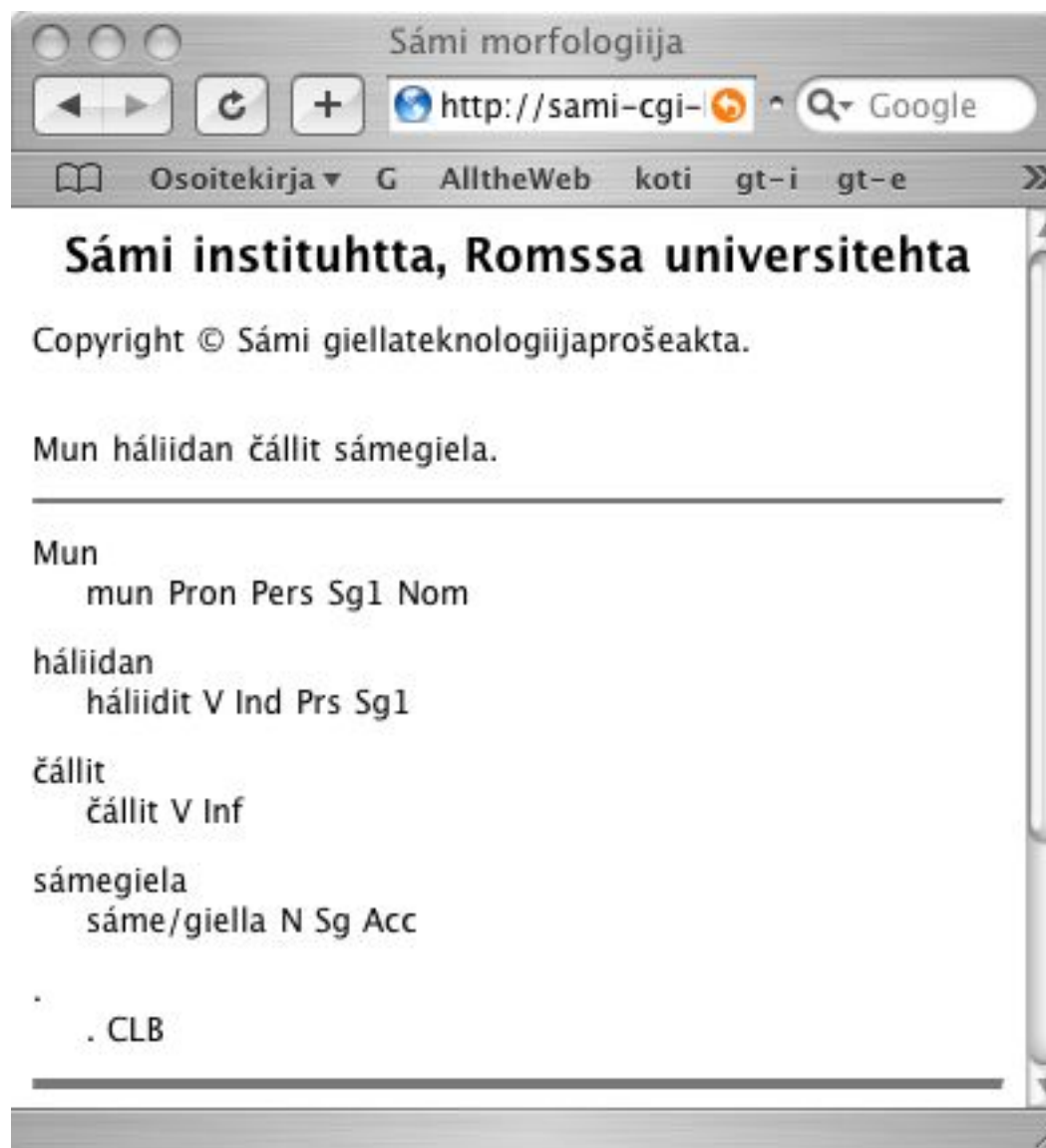
Prográmma ii leat vel gárvvís. Jos fuomášat meattáhusaid dahje boastu-analiissaid, de sáhtát čállit reivve [Sámi giellateknologijaprošektii](#).

The program is under construction. If you notice errors of any kind, please report them to [The Sámi language technology project](#).

Trond Trosterud

Last modified: Thu Dec 18 14:07:45 EET 2003

Svaret som maskina gjev er:



visl

Visual **s**yntax **l**earning

- Bakgrunn: «Danske studentar har ikkje lenger latin som obligatorisk fag på ungdomsskole og gymnas => dei kan ikkje grammatikk»
- Løysning: «Lær dei grammatikk – bruk *visl*»



Games	Sentence Analysis	Quizzes & tests	Corpus search	Languages
-------	-------------------	-----------------	---------------	-----------

Research

- [Constraint Grammar](#)
- [Corpus Linguistics](#)
- [Treebanks](#)
- [Machine Translation](#)
- [Question Answering](#)

Projects

- [PaNoLa](#)
- [Nordic Treebank Network](#)
- [Named Entity Recognition](#)
- [Korpus90/2000](#)
- [Arboretum](#)
- [AC/DC](#)
- [Floresta Sintá\(c\)tica](#)
- [Arboratoire/Freebank](#)



Teaching

- [Grammar Games](#)
- [Quizzes](#)
- [Sentence Analysis](#)
- [Sentence Lab](#)
- [Grammy](#)
- [Lecture Notes](#)

Projects

- [VISL-GYM](#)
- [VISL-SEM](#)
- [Almen Sprogforståelse](#)
- [VISL-Folke](#)
- [GREI](#)

Dear visitor,

The VISL-project is presently carrying out a user-evaluation with the intention of increasing the homepage's usability. We would appreciate if you would participate by filling out a short questionnaire.

Thank you.

[Click here to fill out the questionnaire in english](#)

[No, I do not want to participate.](#)



Liste over språka som er representert
i *visl*-systemet

Alle språk vi underviser i på UiT er
representert

visl-systemet er dermed relevant
også for andre enn samisk

Ordklasseøvingar

Finnish

Help!  

Se että Maa on
pyöreä on jo
tiedetty pitkään

Repeat

New Sentence

Start Select



Well done. You made my day!

Total time: 40.073 seconds

Wrong choices: 0

☒ Sound

☒ Hippo

Norwegian (nynorsk)

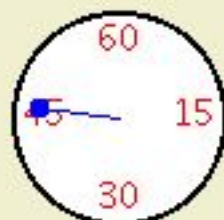
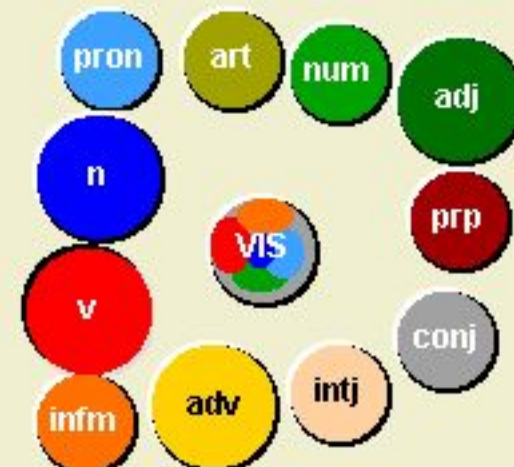
Help!  

Høgtideleg **rekte** dommaren Monica diplomet

Gjenta

Ny setning

Start val



Sånn skal det vere.

Ord som står att: 1

Galne val: 0

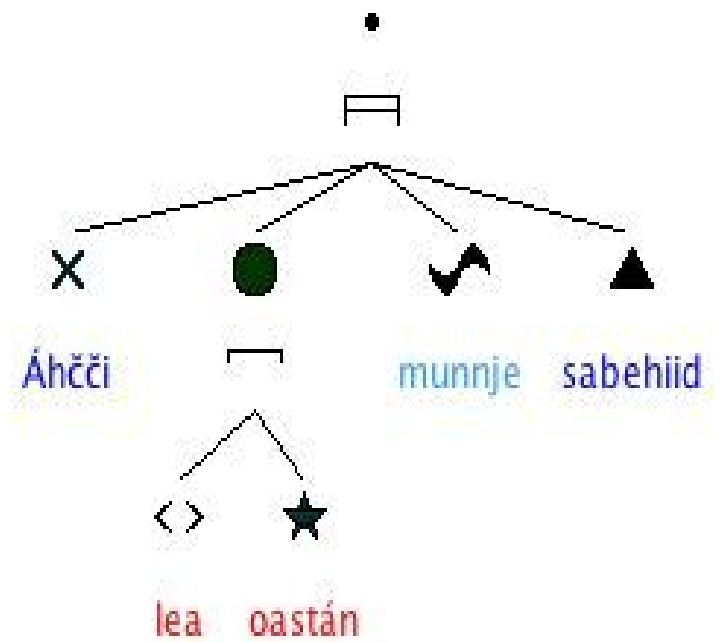
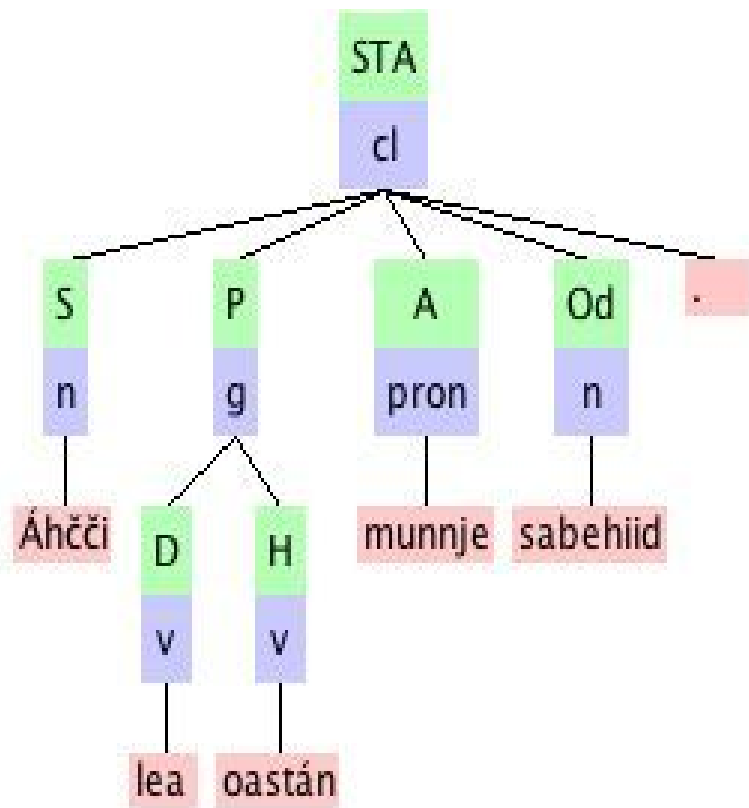
☒ Lyd

☒ Flodhest

Setningsanalyse

"<Áhčči>"
 "áhčči" N Sg Nom @SUBJ
 "<lea>"
 "leat" V Ind Prs Sg3 @+FAUXV
 "<oastán>"
 "oastit" V PrfPrc @-FMAINV
 "<munnje>"
 "mun" Pron Pers Sg1 Ill @ADVL
 "<sabehiid>"
 "sabet" N Pl Acc @OBJ

SOURCE: text
 SME65 Áhčči lea oastán munnje sabehiid.
 A1
 S:n('áhčči',sg,nom) Áhčči
 Vaux:v('leat',ind,pr,3sg) lea
 P:v('oastit',pcp2) oastán
 A:pron('mun',<pers>,1sg,ill) munnje
 Od:n('sabet',pl,acc) sabehiid



"<Jávregáttis>"
 "Jávregáddi" N Prop Sg Loc @ADVL
 "<leat>"
 "leat" V Ind Prs Pl3 @+FAUXV
 "<guokte>"
 "guokte" Num Sg Nom @SUBJ-QH
 "<olbmo>"
 "olmmoš" N Sg Gen @<GQ

SOURCE: Nickel: "Samisk grammatikk"

SME7 Jávregáttis leat guokte olbmo. (Ved elvebredden er det to mennesker.)

A1

STA:cl

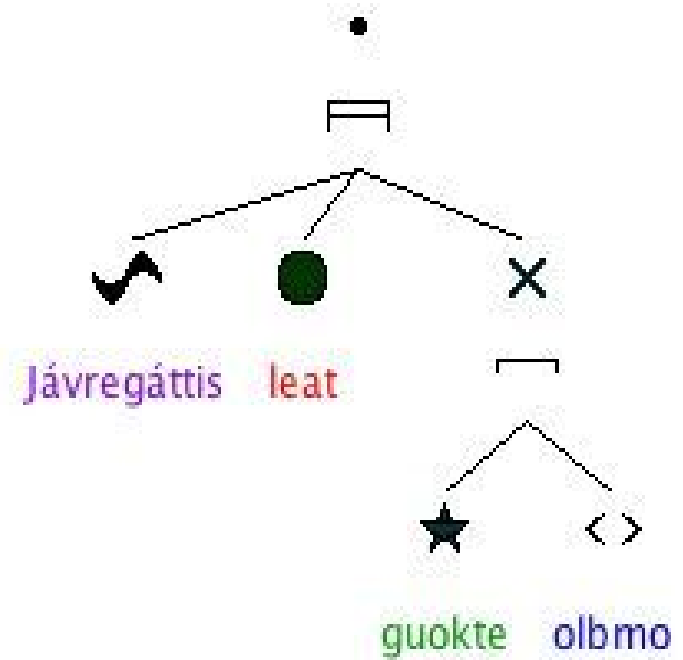
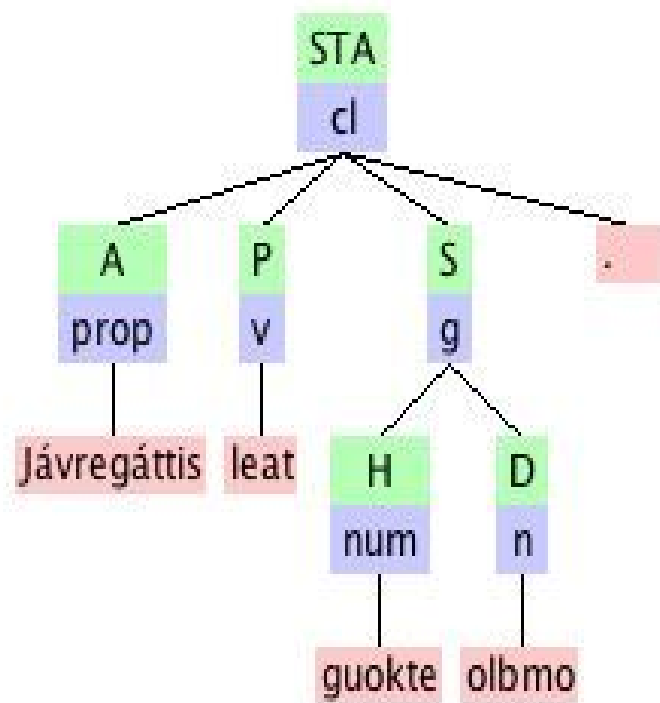
A:prop('Jávregáddi',sg,loc) Jávregáttis

P:v('leat',ind,pr,3pl) leat

S:g

=H:num('guokte',sg,nom) guokte

=D:n('olmmoš',sg,gen) olbmo



SOURCE: MOEDE I PETERSBORG
 GYM90) В магазине они купили пакет молока.
 A1
 STA:cl
 A:g
 =H:prp В
 =D:n(masc,sg,loc) магазине
 S:pron(pers,3pl,nom) они
 P:v(pl,past,pf) купили
 Od:g
 =H:n(masc,sg,acc) пакет
 =D:n(neut,sg,gen) молока



Russian

Help!  

Она там ?

Repeat

New Sentence

Start Select



Superb performance. Are you
a linguist, or what?

Total time: 7.152 seconds

Wrong choices: 0

☐ Sound

☒ Hippo

Visl-filosofi

- Felles infrastruktur for alle språk (27 ulike språk)
- Alle språka har eit sett av analyserte setningar, dette settet blir brukt i mange ulike pedagogiske spel
 - Setningsanalyse (lage syntaktiske trestrukturar)
 - Ordklassespel (ballongferd, labyrint, maleskrin, postkontor, skytespel, romferd, tetris)
 - Øving i morfologisk bøying
 - Estimering av kor stort passivt ordforråd eleven har

Konsekvens for samisk

- Med ein automat kan vi automatisk analysere kjeldekode (står att: program for å generere trestruktur)
 - Dermed har vi automatisk tilgang til ei ti-tal ulike pedagogiske verkty
- Desse verktya er interaktive
 - Dette er eit kvantesprang samanlikna med andre program, der internett eigentleg berre er ein alternativ publikasjonskanal for papirbasert materiale

Oversikt over foredraget

1. Grammatisk analyse

a) morfologisk analyse av samisk

i. morfologisk analyse av einskildord

ii. morfofonologiske justeringsreglar

b) analyse av ord i kontekst (disambiguering)

2. Pedagogiske applikasjonar

3. Modellar for samarbeid om samisk språketeknologi

Samarbeid om pedagogiske program for samisk?

- Samiske pedagogiske miljø
 - Sametingets opplæringsavdeling, <http://www.e-skuvla.no>, Sámi oahpponeahtta, skolane, universiteta, språksentra
- Samiske språkforskingsmiljø:
 - NSI, Sámi allaskuvla, UiT, Uleåborg, Umeå, Helsingfors
- Samiske normative organ
 - Samisk språkavdeling
- Samiske forlag
 - Lærebøker, leksikografiske prosjekt

Potensielt interesserte utviklarar?

- Alle kan vere med og utvikle kjeldekoden vår (open kjeldekode)
- Pedagogisk utvikling
 - Alle dei samiske grammatikklærebøkene kan bli del av dette interaktive pedagogiske programmet
 - Med systematisk klassifisering av ordforrådet kan vi inkludere omsetjingsøvingar, og systematisk ordinnlæring
 - Folk som vil ha ein spesiell versjon av programma, til sitt spesielle bruk (slik det norske Sametinget ville)

Samarbeid om *visl* med andre språkmiljø?

- Visl i Danmark:
 - *Visl* er standard programvare i danske gymnas
 - Kryds- og bolle-systemet vart utarbeidd på 1920-talet
 - Syddansk universitet har integrert *visl* i forprøveundervisninga si
- Visl i Oslo:
 - <http://www.tekstlab.uio.no/grei/>
- Visl i Tromsø:
 - Samisk. Andre?

Oppsummering

- Arbeid i dag:
 - **Tromsøprosjektet**: Analysator, disambiguator for nord- og lulesamisk, og korpus med søkbart grensesnitt
 - **Sametinget**: ordretteprogram for nord- og lulesamisk
 - **Pedagogisk programvare** i samarbeid med *visl*
 - **Overføring av infrastruktur** til analysatorar for andre uralske språk (grønlandsk, komi, marisk, ...)

Framtidsperspektiv

- Pedagogisk programvare
 - Vi håpar dei blir tatt i bruk, og dermed også vidareutvikla
 - Vi vil gjerne ha fleire tekstar, og fleire samiske språk
- Grammatisk analysator
 - Først må dei bli betre, og vi må få fleire tekstar
 - Seinare vil vi inkludere semantikk
- Skrivestøtteverktøy og administrativ programvare
 - Etter stavekontroll vil vi ha grammatikkontroll
 - Vi vil ha fleirspråkleg, grammatisk basert informasjonssøk, og terminologisk arbeid basert på parallellkorpora

Alt dette er fritt tilgjengeleg for alle