Samisk disambiguering 2005

Saara Huhmarniemi
Marit Julien
Ilona Kivinen
Pekka Sammallahti
Trond Trosterud
Linda Wiechetek
http://giellatekno.uit.no/

Tema

1. Prosjektet vårt
2. Samarbeidsprosjekt
3. Samarbeid
4. Framdrift
5. Resultat
6. Resten av prosjektperioden, evaluering
7. Framtidsvyer ut over prosjektperioden

1. Prosjektet vårt

- *Pekka Sammallahti* – prosjektleiar
- *Trond Trosterud* – arbeider heile perioden
- *Marit Julien*
  - arbeidde med disambigueraren i 14 mnd (1.9.04 – 1.11.05)
- *Linda Wiechetek* – tilsett i 80 % stilling frå 15.10
  - arbeider med disambigueraren
- *Ilona Kivinen* – arbeider på timebasis ~ 20 %
  - arbeider med leksikon
- *Saara Huhmarniemi* (svangerskapsperm 15.10.04 – 31.12.05)
  - programmerar (50 %)

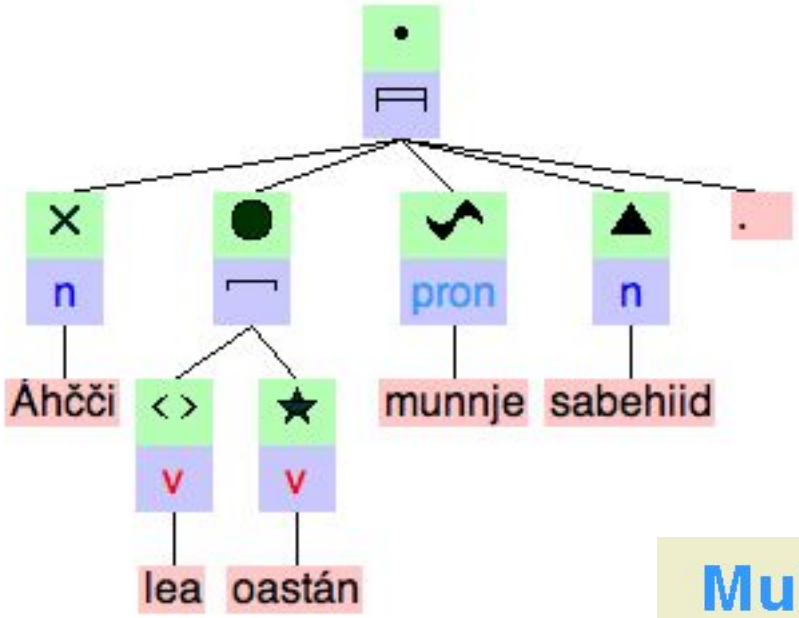
Mål:

- Lage morfologisk disambiguator for *nordsamisk*
- Lage analysator og grunnleggjande disambiguator for *lulesamisk*
- Samle tekstar til eit *korpus*
- Lage *grafisk brukargrensesnitt* for det tagga korpuset

2. Samarbeidsprosjekt

- Sametingets retteprogramprosjekt
- PaNoLa+ pedagogiske program for små nordiske språk

PaNoLa+



Korpus på 225 setningar, automatisk analysert og manuelt korrigert

Sametinget sitt retteprogramprosjekt

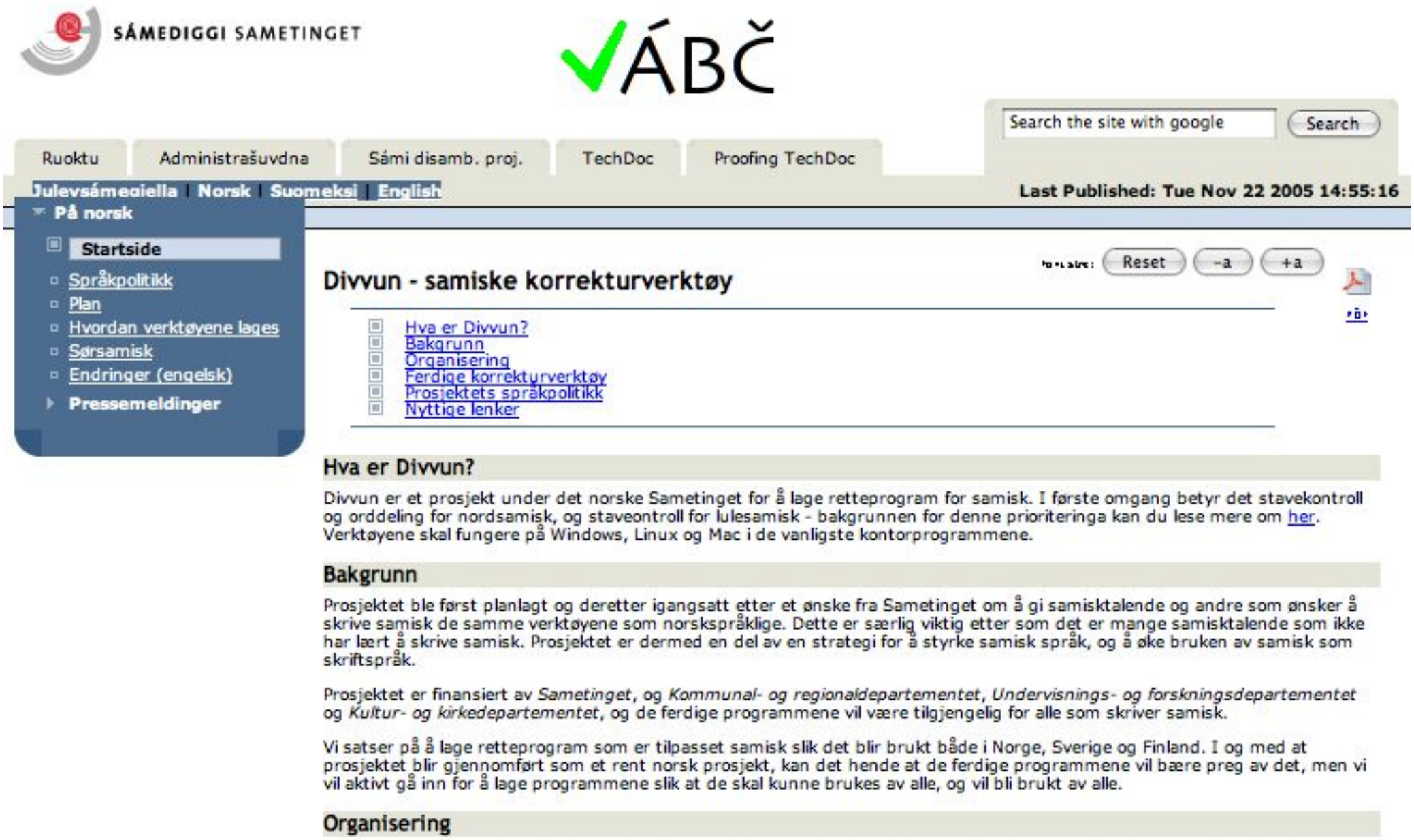
- Prosjektperiode: 2004 – 2007
- Stavekontroll for nord- og lulesamisk
- Folk
    - *Sjur Moshagen*, prosjektleiar (Helsingfors)
    - *Børre Gaup*, programmering / korpus (Tromsø/Kautokeino)
    - *Thomas Omma*, lingvistikk (Tromsø)
    - *Maaren Palismaa*, lingvistikk (50 %) (Kautokeino)
    - *Tomi Pieski*, programmering (Tromsø)

Samarbeid

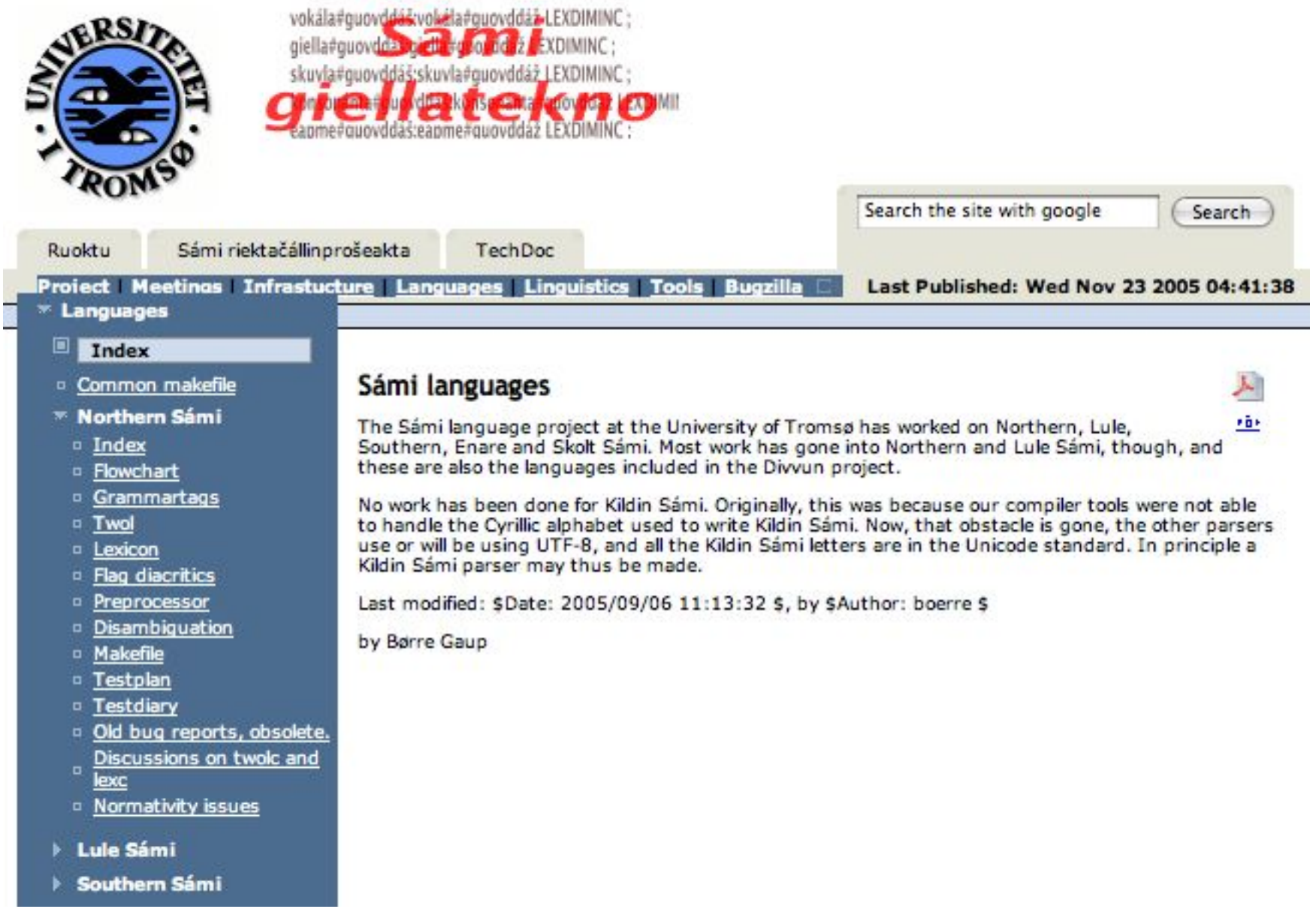
- Fellesarbeid:
  - morfofonologi, morfologi og leksikon
  - korpusinnsamling og -infrastruktur
  - dokumentasjon
  - feildatabase
- Separat arbeid:
  - *UiTø*: Disambiguering, grafisk korpusgrensesnitt, deskriptiv parsar
  - *Sametinget*: Korrekturprogram integrert i programvare, normativ parsar

Dokumentasjon: *http://giellatekno.uit.no/*

Dokumentasjon: *http://divvun.no/*



Felles teknisk dokumentasjon:



Feildatabase: *http://129.242.176.176/giellatekno/bugzilla/*

Bugzilla Version 2.20

Bug List

Sun Dec 4 2005 10:28:07

Café: plaats waar mensen die niets te vertellen hebben mensen die niet kunnen luisteren aan flarden praten.

17 bugs found.

ID	Opened	Changed	Assignee	Reporter	Status	Product	Comp	Summary
6	2004-12-28	2005-09-11	Tomi Pieski	Trond Trosterud	NEW	sme lexi	Numerals	Num tag is needed in compounds, but stripped in lookup2cg
56	2005-03-04	2005-11-19	Trond Trosterud	Lena Gaup	ASSI	sme lexi	Continua	-headdjiid and -heddjiid
77	2005-05-11	2005-10-06	Trond Trosterud	Thomas Omma	ASSI	sme morp	Rule com	consonantchange in the end of verbstem
158	2005-07-11	2005-09-11	Trond Trosterud	Ilona Kivinen	ASSI	sme lexi	Numerals	Num+Sg+Gen+logi
169	2005-08-04	2005-09-11	Trond Trosterud	Ilona Kivinen	NEW	sme lexi	Numerals	golbmalohkása
176	2005-08-18	2005-09-11	Trond Trosterud	Ilona Kivinen	NEW	sme lexi	Numerals	beal+Ord
186	2005-09-12	2005-10-17	Trond Trosterud	Thomas Omma	ASSI	sme morp	Rule com	No dipht. simpl in actor nouns before uj
193	2005-09-29	2005-11-11	Trond Trosterud	Thomas Omma	ASSI	smj morp	Rule com	oa->å dipht. simpl. in actor nouns
198	2005-10-22	2005-11-01	Tomi Pieski	Trond Trosterud	ASSI	Corpus	Text cor	xsl script for Bible files does not single out chapter he...
201	2005-10-31	2005-11-09	Børre Gaup	Trond Trosterud	NEW	Infrastr	Localisa	We have an UTF-8 problem with our web pages.
209	2005-11-15	2005-11-15	Trond Trosterud	Trond Trosterud	NEW	sme lexi	Numerals	Hyphen count as numeral of its own.
211	2005-11-17	2005-11-21	Tomi Pieski	Trond Trosterud	NEW	Infrastr	Localisa	Perl problems with UTF-8
214	2005-11-21	Fri 13:45	Tomi Pieski	Trond Trosterud	NEW	Pre- and	catxml	Paragraphs failing to be identified as such.
223	Mon 19:48	Mon 19:48	Børre Gaup	Trond Trosterud	NEW	Corpus	xsl conv	Errors in the conversion.
225	Thu 23:10	Fri 13:38	Børre Gaup	Trond Trosterud	NEW	Corpus	xsl conv	Some of the converted corpus files are broken
226	Fri 00:17	Fri 10:59	Tomi Pieski	Trond Trosterud	NEW	Infrastr	Compilat	bin/missing is not generated
227	Sat 10:39	Sat 10:52	Børre Gaup	Trond Trosterud	NEW	Corpus	xsl conv	All the files in gt/sme/admin/depts/ (save one, the fakta...

17 bugs found.

Long Format | CSV | RSS | iCalendar | Change Columns | Change Several Bugs at Once | Edit Search | Remember search as

CVS

```
revision 1.87
date: 2005/11/08 10:16:53;  author: trond;  state: Exp;  lines: +13 -16
Changed both the Sets and Definitions section, in order to fix the
issues raised in the #193 bug.

The Cns1 was split in a Cns1 that cannot be the first part of a
two-letter G3, and a Cns2 without this restriction. In the definitions
section, two sets LCnsPhon1 and LCnsPhon were difined accordingly.

Then the LowerG2 as redefined. The two-consonant string only included
the narrower LCnsPhon1 set, and the G2 sequences that contained the
consonants b, d, g, k as first consonants were explicitly listed (the
dj, dn, dnj, gŋ, kn, kk, bm, bd, bj, bl, bn, bnj, br, bgs, btj,
bts. All this in order NOT to include the two-consonant G3s, namely
dts, dtj, bb, dd, gg, ks, kt, ktj, kts.

The 3-consonant G2 were fixed in the previous update.
----------------------------
revision 1.86
date: 2005/11/07 14:00:28;  author: thomas;  state: Exp;  lines: +4 -4
Added l3 and n3 on line 23. Changed to double pointing arrow in rules for deleting l3 and n3:
"Deleting Final l3 in Short Essive of Uneven Syllables"
 l3:0 <=> Vow: _ Q1: X1: n ;
Changed right context in rule for "Compulsatory lengthening in grade I even-syllables" from (Cns)
to (Cns:0 | Cns:):
  a:á<=> (Cns) [ a|i|u ] [ Cns:0 Cns: | Cns: Cns:0 ] _ (Cns:0 | Cns:) WeG: ; ! in grade I
----------------------------
revision 1.85
date: 2005/11/07 12:58:50;  author: thomas;  state: Exp;  lines: +16 -22
Adjusted the rule-set on æ:e dihpt. simpl. according to our G3 revision:
 æ:e  => [ #: | Cns: | Dummy: ] _ LowerG12 o: [ (Y6:) X4: | [ Y6: | Q1: ] X5: ] j ; !6 lll&ComSg.
!instead of !line
         [ #: | Cns: | Dummy: ] _ LowerG12 o: (Cns:*) [ Y1: | Y2: | Y3: | Y4: | Y5: | Y7: | Y9: |
X7: ] ; !7 Long pass, Dim and Prs Sg3 etc.
:
```

Diskusjonsliste

Subject	From	Date
Re: corpus.dtd	Sjur	21.11.2005
Re: corpus.dtd	Saara Huhmarniemi	21.11.2005
Re: corpus.dtd	Trond Trosterud	22.11.2005
▶ A new name lexicon	Trond Trosterud	23.11.2005
▼ north sami compounding	Thomas	19.11.2005
Re: north sami compounding	Thomas	21.11.2005
Re: north sami compounding	Sjur	21.11.2005
Re: north sami compounding	Thomas	21.11.2005
Re: north sami compounding	Trond Trosterud	21.11.2005
Re: north sami compounding	Sjur	21.11.2005
Re: north sami compounding	Maaren Palismaa	23.11.2005
Re: north sami compounding	Thomas	24.11.2005
Re: north sami compounding	Maaren Palismaa	23.11.2005
lookup2cg	Saara Huhmarniemi	25.11.2005
web address regex in gt/common	Trond Trosterud	28.11.2005
▼ corpus: xsl files under version control	Saara Huhmarniemi	25.11.2005
Re: corpus: xsl files under version control	Trond Trosterud	01.12.2005
▼ Kvensk place name project	Sjur	23.11.2005
Re: Kvensk place name project	Trond Trosterud	01.12.2005

4. Framdrift

- Nordsamisk
- Lulesamisk
- Korpusgrensesnitt

Nordsamisk

- Ein fungerande versjon av disambiguatoren er ferdig
  - Opne lingvistiske spørsmål
    - MF: Fakultativ monoftongisering, selektiv utlydsherding...
    - Leksikon: Leksikalske luker, forbetra namneleksikon
    - Disamb: Akk/Gen, Kom Sg/Lok Pl, leksikalsk homonymi, i det heile meir arbeid med disambigueringsreglane
  - Meir analysetekniske spørsmål:
    - Analyse av talord, forkortingar, ikkjespråklege tekstelement (webadresser o.l.), ...
    - I ein del tilfelle er det ope korleis vi skal tagge

```
"<Amma>"
        "amma" Pcle @PCLE
"<mii>"
        "mun" Pron Pers Pl1 Nom @SUBJ
"<eat>"
        "ii" V IV Neg Ind Pl1 @+FAUXV
"<leat>"
        "leat" V IV Ind Prs ConNeg @-FAUXV
"<máksán>"
        "máksit" V TV PrfPrc @-FMAINV
"<?>"
        "?" CLB <<<
"<De>"
        "de" Adv @ADVL
"<leat>"
        "leat" V IV Ind Prs Sg2 @+FAUXV
        "leat" V IV Ind Prs Pl1 @+FAUXV
        "leat" V IV Ind Prs Pl3 @+FAUXV
"<máksán>"
        "máksit" V TV PrfPrc @-FMAINV
"<.>"
        "." CLB <<<
"<Mun>"
        "mun" Pron Pers Sg1 Nom @SUBJ
"<máksen>"
        "máksit" V TV Ind Prt Sg1 @+FMAINV
"<100>"
        "100" Num Acc @QN>
"<ruvnno>"
        "ruvdno" N Sg Acc @OBJ
"<duvle>"
        "duvle" Adv @ADVL
"<.>"
        "." CLB <<<
```

```
"<Geahča>"
        "geahčča" V TV Imprt Prs Sg2 @+FMAINV
"<,>"
        "," CLB
"<nieida>"
        "nieida" N Sg Nom @SUBJ
"<šaddá>"
        "šaddat" V IV Ind Prs Sg3 @+FAUXV
"<máná>"
        "mánná" N Sg Gen @GP>
"<vuostái>"
        "vuostái" Po @ADVL
"<ja>"
        "ja" CC @CC
"<riegádahttá>"
        "riegádahttit" V TV Ind Prs Sg3 @+FMAINV
"<bártni>"
        "bárdni" N Sg Acc @OBJ
"<,>"
        "," CLB
"<ja>"
        "ja" CC @CC
"<son>"
        "son" Pron Pers Sg3 Nom @SUBJ
"<gohčoduvvo>"
        "gohčodit" V TV Pass Ind Prs Sg3 @+FMAINV
"<Immanuelin>"
        "Immanuel" N Prop Mal Ess @SPRED
"<->"
        "" Num @NUMBER
        "-" PUNCT
"<dat>"
        "dat" Pron Pers Sg3 Nom @SUBJ
"<mearkkaša>"
        "mearkkašit" V TV Ind Prs Sg3 @+FMAINV
"<:>"
        ":" CLB
"<Ipmil>"
        "ipmil" N Sg Nom @SUBJ
"<lea>"
        "leat" V IV Ind Prs Sg3 @+FAUXV
"<minguin>"
        "mun" Pron Pers Pl1 Com @ADVL
"<.>"
        "." CLB <<<
```

```
"<Grete>"
        "Grete" N Prop Fem Sg Acc @OBJ
        "Grete" N Prop Fem Sg Nom @SUBJ
        "Grete" N Prop Fem Sg Nom @SPRED
        "Grete" N Prop Fem Sg Acc @OPRED
        "Grete" N Prop Fem Sg Attr @PROP>
        "Grete" N Prop Fem Sg Gen @ADVL
"<vujii>"
        "vuodjit" V TV Ind Prt Sg3 @+FMAINV
"<Kari>"
        "Kari" N Prop Fem Sg Acc @OBJ
        "Kari" N Prop Mal Sg Acc @OBJ
        "Kari" N Prop Fem Sg Nom @SUBJ
        "Kari" N Prop Mal Sg Nom @SUBJ
        "Kari" N Prop Mal Sg Nom @SPRED
        "Kari" N Prop Fem Sg Gen @GN>
        "Kari" N Prop Fem Sg Nom @SPRED
        "Kari" N Prop Mal Sg Gen @GN>
        "Kari" N Prop Fem Sg Attr @PROP>
        "Kari" N Prop Mal Sg Acc @OPRED
        "Kari" N Prop Fem Sg Acc @OPRED
        "Kari" N Prop Mal Sg Attr @PROP>
"<biilla>"
        "biila" N Sg Acc @OBJ
        "biila" N Sg Gen @GN>
        "biila" N Sg Acc @OPRED
"<Jotunheimena>"
        "Jotunheimen" N Prop Plc Sg Acc @OBJ
        "Jotunheimen" N Prop Plc Sg Gen @GP>
        "Jotunheimen" N Prop Plc Sg Acc @OPRED
"<ćađa>"
        "ćađđa" N Sg Acc @OBJ
        "ćađđa" N Sg Gen @ADVL
        "ćađđa" N Sg Acc @OPRED
        "ćađa" Pr @ADVL
        "ćađa" Po @ADVL
        "ćađa" Adv @ADVL
"<.>"
        "." CLB <<<
```

```
"<Grete>"
        "Grete" N Prop Fem Sg Nom @SUBJ
"<vujii>"
        "vuodjit" V TV Ind Prt Sg3 @+FMAINV
"<Kari>"
        "Kari" N Prop Fem Sg Gen @GN>
        "Kari" N Prop Mal Sg Gen @GN>
"<biilla>"
        "biila" N Sg Acc @OBJ
"<Jotunheimena>"
        "Jotunheimen" N Prop Plc Sg Gen @GP>
"<ćađa>"
        "ćađa" Po @ADVL
"<.>"
        "." CLB <<<
```

Testresultat desember 05

- Kjenne att ordformer:
  - Recall-token: 96.9 %
  - Recall-type: 89,2 %
- Morfologisk og syntaktisk disambiguering:
  - Recall: 98 % - 99 %
  - Presisjon:
    - a. 93 % - 94 % (utan kasusmerka Num og Abbr)
    - b. 61 % - 62 % (kasusmerka Num og Abbr)

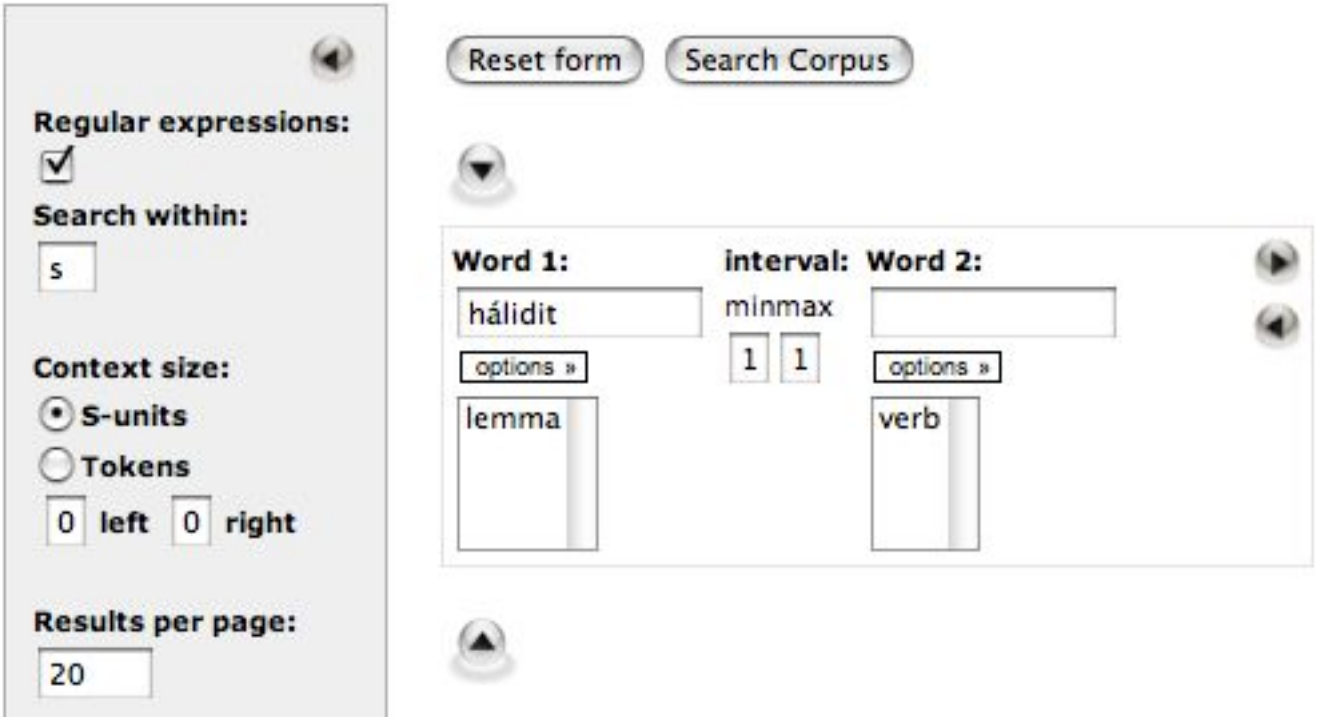
Lulesamisk

- Morfologi
  - Positive effektar av å arbeide parallelt med lule- og nordsamisk morfofonologi
  - Samarbeid med Sametingsprosjektet om mf-reglar
- Leksikon
  - Problem: Framleis ikkje noko leksikon
    - etter kvart spøker det for den lulesamiske disambiguatoren...
- Disambiguator
  - ... ikkje starta opp, kan framleis bli nyttig referanse til nordsamisk

Korpus

- I samarbeid med Sametinget:
  - Vi har prosedyrer for xml-handsaming av korpus
    - Dette er arbeidskrevjande, mange opne spørsmål
  - Vi forhandlar framleis med dei viktigaste tekstprodusentane
    - Byråkratisk tekst ok, avistekst *i prinsippet* ok, skjønnlitteratur er framleis eit ope spørsmål
- I samarbeid med Tekstlaboratoriet i Oslo:
  - Arbeid med brukargrensesnitt

Betaversjon av grensesnitt



Resultat

Query string: "[lemma="h\'alidit" %c] []{1,1} [(POS="v")] within s"
Action: Count Download Sort Collocations Annotate
Hits found: 16
Results pages:
1

1198 Beassášmállásat Suvrrokeahtes láibbiid basiid vuosttaš beaivvi máhttájeaddjit bohte Jesusa lusa ja jerre : " Gosa don **hálidat min lágidit** alccesat beassášmállásiid ?

1729 Doppe son manai guossin muhtun dállui ii ge **hálidan ovttage diehtit** dan , muhto dat ii lean čiegadeames .

2119 Beassášmállásat ráhkaduvvojit Suvrrokeahtes láibbiid basiid vuosttaš beaivvi , go beassášláppis njuvvojuvvui , máhttájeaddjit jerre sus : " Gosa don **hálidat min mannat** ráhkadit dutnje beassášmállásiid ?

3002 Dii dáhttubehtet čohkkát synagogaid buoremus sajiin ja **hálidehpet olbmuid buorástahttit** din márkanšiljus .

3380 Son **hálidii sakka oaidnit** makkár olmmoš Jesus lea , muhto lei uhcci šattus ii ge beassan su lahka olbmuid dihtii .

3649 Pilatus sárdnugođii fas sidjiide , dasgo son **hálidii áinnas luoitit** Jesusa .

5797 Go sii gulle Bávlosa sárdnumin jábmiid bajášcuožžileami birra , de muhtumat bilkidedje su , muhto earát fas dadje : " Mii **hálidivččiimet áinnas gullat** du sárdnumin dán birra nuppe háve ge .

5908 Nu son lei mearridan , dasgo son **hálidii mannat vácci** dohko .

6278 Muhto mii **hálidit áinnas gullat** maid don oaivvildat , dasgo mii diehtit ahte olbmot vuostálastet dán searvvi juohke báikkis .

Framdrift - oppsummering

		Start	Slutt	Revidert
1	Språkuavhengig preprosessering	2004-1	2004-1	ok
2	Infrastruktur for disambiguering	2004-1	2004-2	ok
3	Korpusgrensesnitt - prototyp	2004-1	2004-4	2005-2
4	Grunnarbeid for nordsamisk	2004-1	2004-4	ok
5	Nordsamisk disambiguering - prototyp	2004-1	2005-2	ok
6	Revidere morfologiske analyseprogram	2004-1	2006-4	ok
7	Grunnarbeid for lulesamisk	2004-3	2005-4	2006-1
8	Lulesamisk disambiguering - prototyp	2004-4	2005-4	2006-2?
9	Parallelltekstkorpora - prototyp	2005-1	2005-2	2006-1
10	Korpusgrensesnitt - beta	2005-1	2005-4	ok
11	Nordsamisk disambiguering - beta	2005-3	2005-4	ok
12	Parallelltekstkorpora - beta	2005-3	2006-1	2006-2
13	Lulesamisk disambiguering - ferdig	2005-4	2006-2	2006-?
14	Nordsamisk disambiguering - ferdig	2006-1	2006-4	ok
15	Korpusgransesnitt - ferdig	2006-1	2006-4	ok
16	Parallelltekstkorpora - ferdig	2006-2	2006-4	ok

5. Resultat

- I dag
  - Vi har avdekka manglar ved referansegrammatikkane, og delvis vore i stand til å komplettere dei
  - Vi har gjort det mogleg å planleggje og gjennomføre andre språkteknologiske prosjekt for samisk
  - Vi har laga ein open infrastruktur som andre ikkje-kommersielle prosjekt er interessert i å bruke (komi, vestgrønlandsk, færøysk, ...)
- Etter prosjektet:
  - Vi har lagt fundamentet for all framtidig samisk språkteknologi
  - Vi har gjort analysert samisk tekst tilgjengeleg for forsking

6. Resten av prosjektperioden

- Lingvistisk arbeid:
  - Analysator for lulesamisk i samarbeid med Sametinget
  - Disambiguator, her med vekt på nordsamisk, evt. arbeid med lulesamisk i den grad det vil vere mogleg
- Korpusarbeidet:
  - Nordsamisk og lulesamisk tekstinnsamling skjer parallelt, i samarbeid med Sametinget sitt prosjekt
  - Det grafiske grensesnittet blir gjort ferdig
- Prioritering:
  - Perfeksjonere den nordsamiske disambiguatoren, og å utnytte samarbeidet med Sametinget (arbeide i lag med dei)

Evaluering

- Evaluering av lingvistiske resultat:
  - Testsett:
    - integrerte testsett for mf-reglar
    - i prinsippet Correct!-korpus for disamb, men bug i vislcg
  - Standardmålingar: Recall, presisjon, ambiguitet
  - Vi planlegg også eit seminar for samiske lingvistar neste år der vi diskuterer metodane og resultata våre
- Evaluering av prosjektplanlegging og samarbeid:
  - Ei ekstern gruppe ved Hum.fak. gjennomfører ei slik evaluering no

Publikasjonar

- 2005
  - 2 internasjonale artiklar i bok (m/fagfelleevaluering)
  - 1 artikkel i publikasjon frå internasjonal konferanse (m/fagfelleevaluering)
- 2006
  - 1 artikkel i publikasjon frå internasjonal konferanse (m/fagfelleevaluering) (til no)

Framtidsvyer

1. Forbetre dei analysatorane vi har
2. Fleire språk: Sørsamisk, andre samiske språk
3. Tekst-til-tale, samarbeid med relevante miljø
4. Ein språkbank for samisk, i samarbeid med norske og samiske miljø
5. Intelligent informasjonssøk

   søk på tvers av ordformer, synonym, språk
6. Intelligente elektroniske ordbøker

   ordbøker som analyserer og genererer ordformer
7. Integrering av analysatorar i pedagogisk programvare

...