Documenting and revitalising the Sámi languages - experiences from written language processing

Lene Antonsen, Saara Huhmarniemi, Trond Trosterud

UNIVERSITETET I TROMSØ

and

Børre Gaup, Sjur N. Moshagen

SÁMEDIGGI SAMETINGET

October 8, 2008

Joint work between UiT and Sámediggi

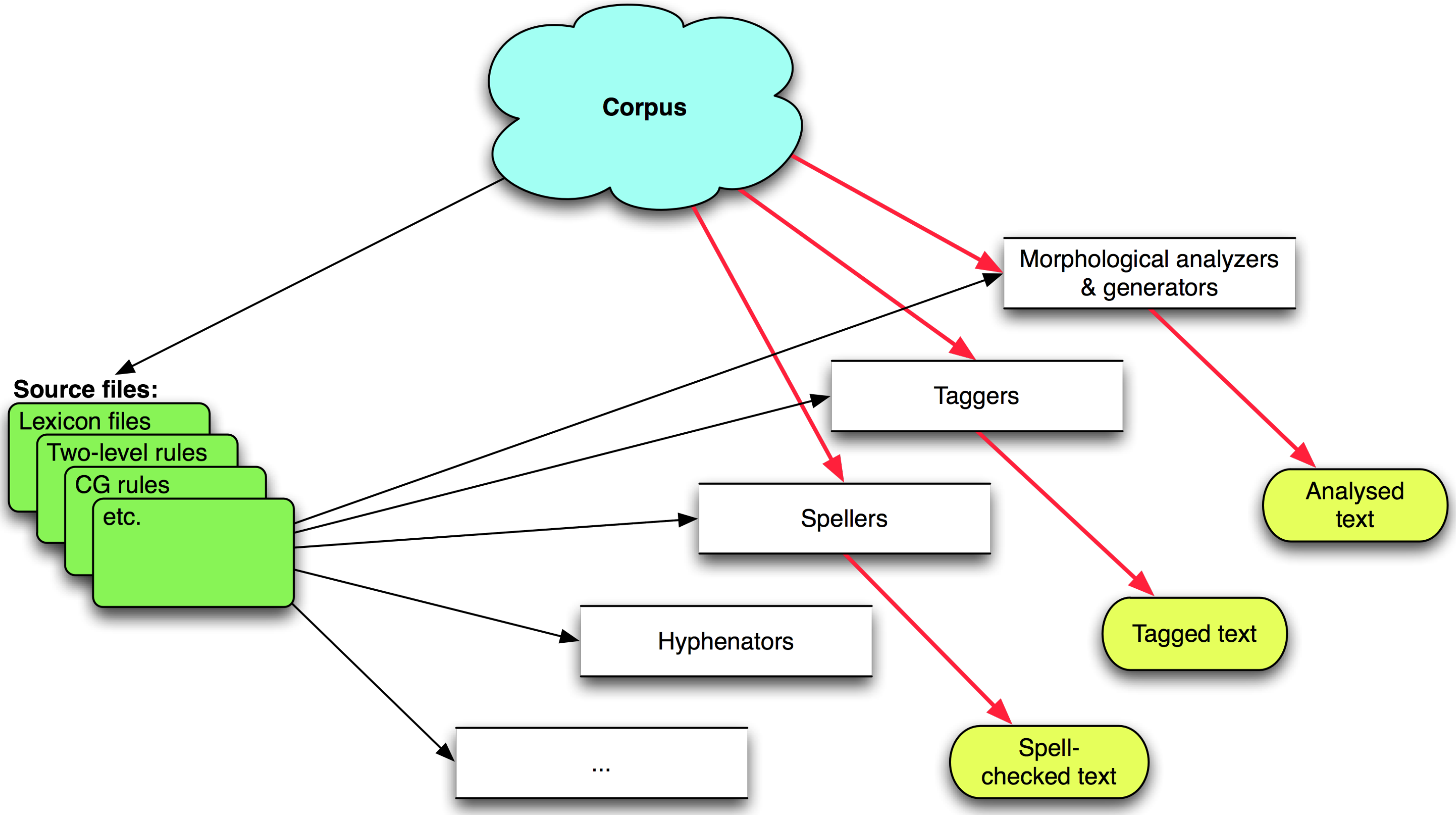
Sharing:

- infrastructure
- linguistic resources
- computer resources
- even man-power to some degree

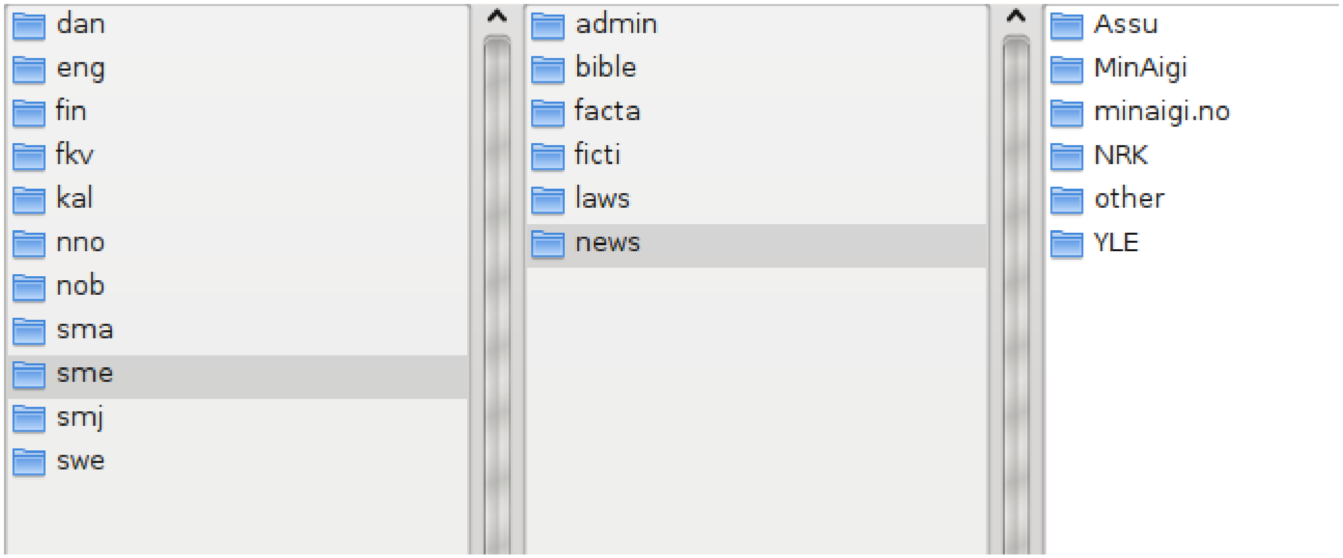
Languages

- We focus on these languages: *North, Lule, South Sámi*
- We have also worked with: *Greenlandic, Faroese, Iñupiaq, Kven, Meänkieli, Komi*
- We have looked at: *Skolt, Inari, Kildin Sámi, Inuktitut*

Overview


Corpus
Source files:
Lexicon files
Two-level rules
CG rules
etc.
Morphological analyzers & generators
Taggers
Spellers
Hyphenators
...
Analysed text
Tagged text
Spell-checked text

Corpus infrastructure - text hierarchy

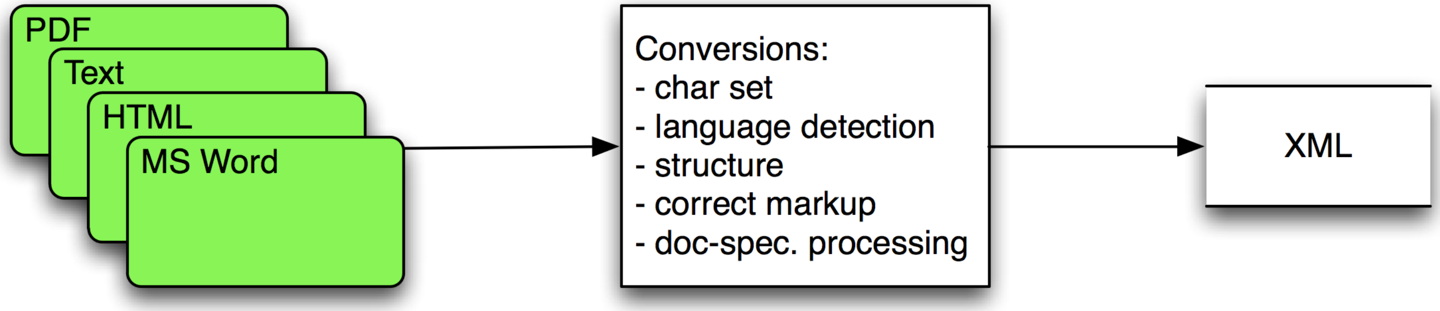


dan
eng
fin
fkv
kal
nno
nob
sma
sme
smj
swe

admin
bible
facta
ficti
laws
news

Assu
MinAigi
minaigi.no
NRK
other
YLE

Corpus infrastructure



PDF

Text

HTML

MS Word

Conversions:
- char set
- language detection
- structure
- correct markup
- doc-spec. processing

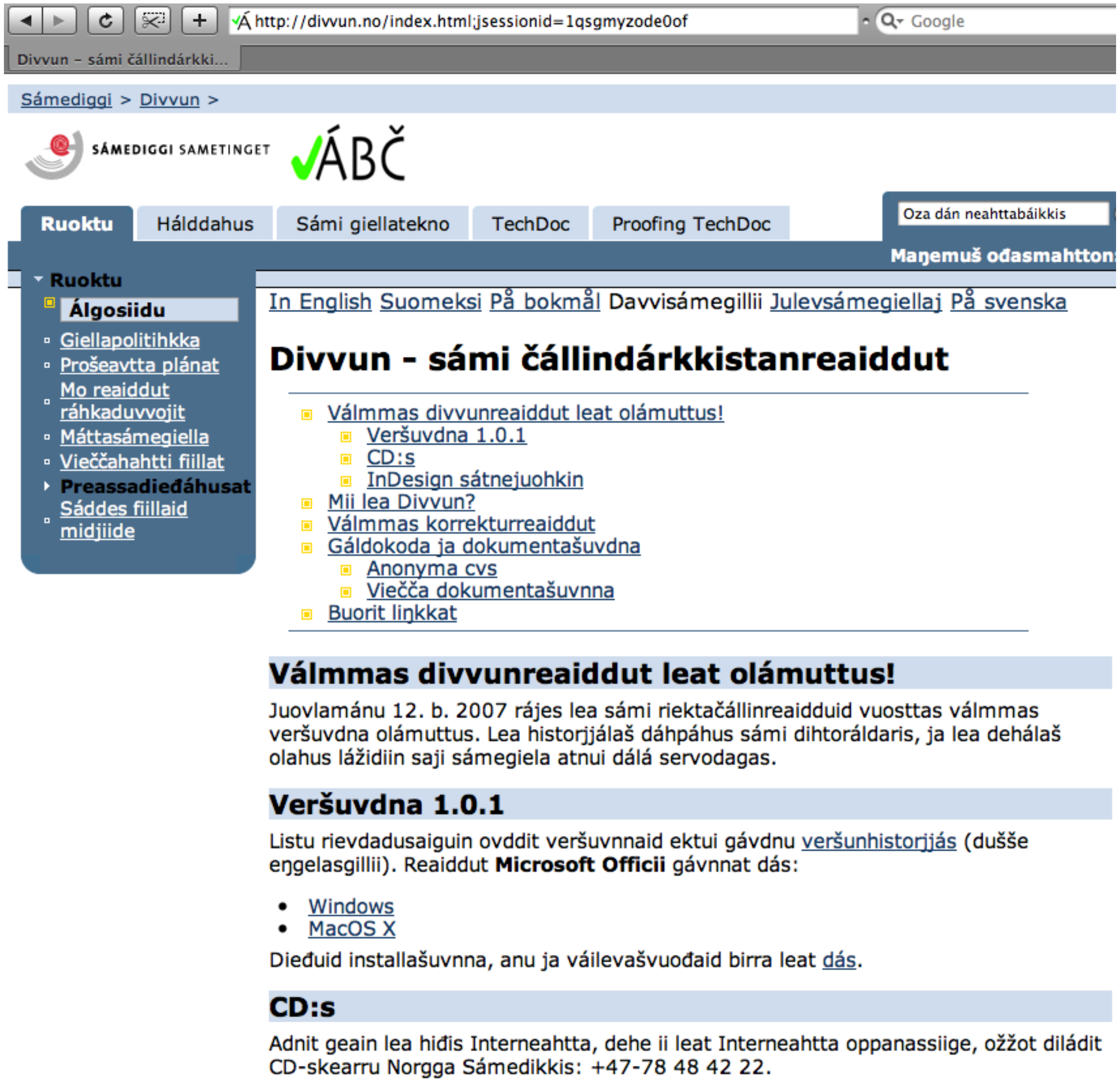
XML

Corpus content overview

Table 1: Number of words in our corpus

Language	North Sámi	Lule Sámi	South Sámi
Admin	2 102 120	148 004	8 749
Bible	202 546	120 287	0
News	4 796 352	7 422	0
Fiction	228 766	12 072	2 025
All words	7 329 784	287 785	10 774

Documentation Infrastructure



Divvun - sámi čállindárkkistanreaiddut

- Válmmas divvunreaiddut leat olámuttus!
  - Veršuvdna 1.0.1
  - CD:s
  - InDesign sátnejuohkin
- Mii lea Divvun?
- Válmmas korrekturreaiddut
- Gáldokoda ja dokumentašuvdna
  - Anonyma cvs
  - Viečča dokumentašuvnna
- Buorit liŋkkat

Válmmas divvunreaiddut leat olámuttus!

Juovlamánu 12. b. 2007 rájes lea sámi riektačállinreaidduid vuosttas válmmas veršuvdna olámuttus. Lea historjjálaš dáhpáhus sámi dihtoráldaris, ja lea dehálaš olahus lážidiin saji sámegiela atnui dálá servodagas.

Veršuvdna 1.0.1

Listu rievdadusaiguin ovddit veršuvnnaid ektui gávdnu veršunhistorjjás (duššeeŋgelasgillii). Reaiddut **Microsoft Officii** gávnnat dás:

- Windows
- MacOS X

Dieđuid installašuvnna, anu ja váilevašvuođaid birra leat dás.

CD:s

Adnit geain lea hiđis Interneahtta, dehe ii leat Interneahtta oppanassiige, ožžot diládit CD-skearru Norgga Sámedikkis: +47-78 48 42 22.

Trond

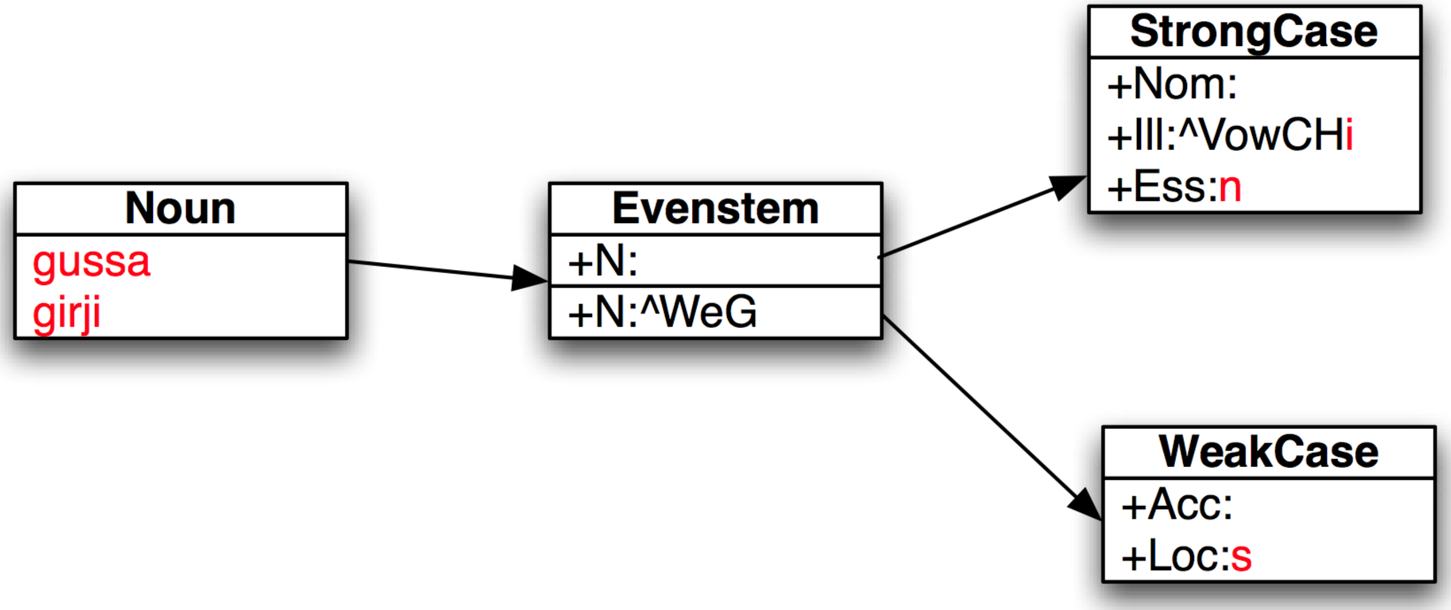
Basic tools

- Morphological analysers / generators
- Morphological disambiguators
- Syntactic analysers

Morphological analysers / generators

- Manually written finite state transducers
- → see grammar as some sort of Red Cross coin automaton
  - (X is a word in the language if there is a path through the automaton which gives X)

Lexical transducer



Noun
gussa
girji

Evenstem
+N:
+N:^WeG

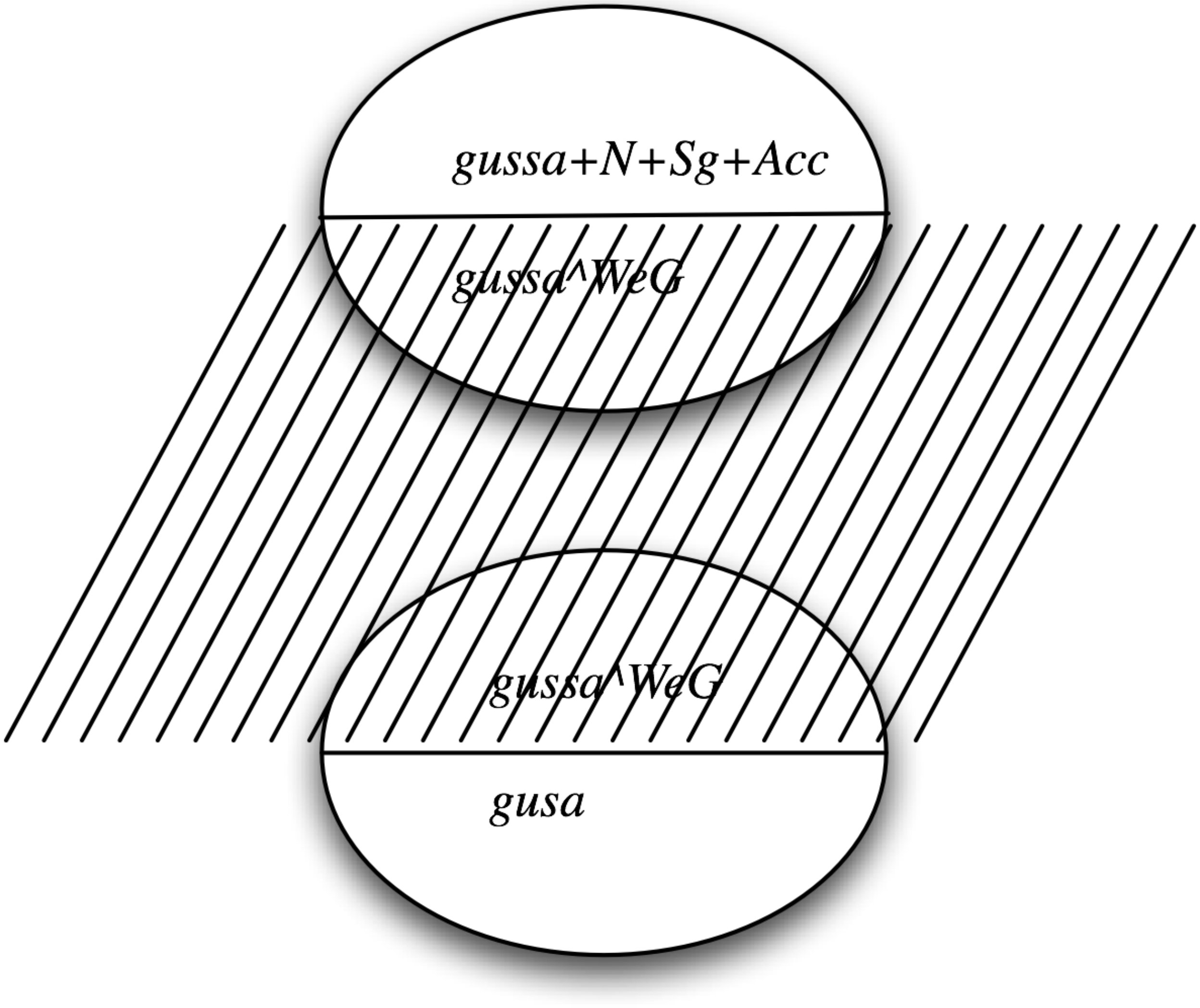
StrongCase
+Nom:
+Ill:^VowCHi
+Ess:n

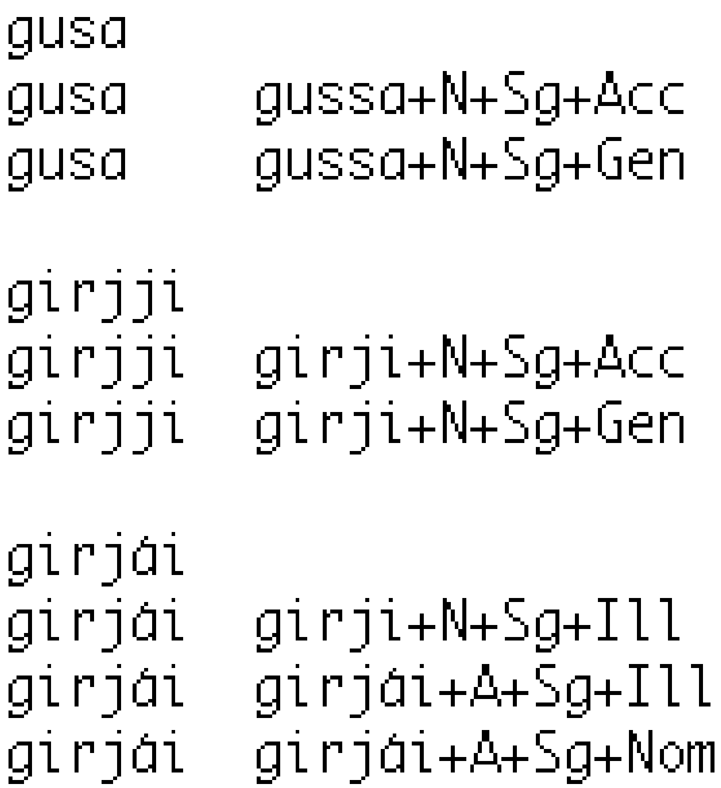
WeakCase
+Acc:
+Loc:s

Phonological transducer

```
ss → s, rj → rjj, ... || _ Vow* WeG ;

i → á || _ VowCH ;
```

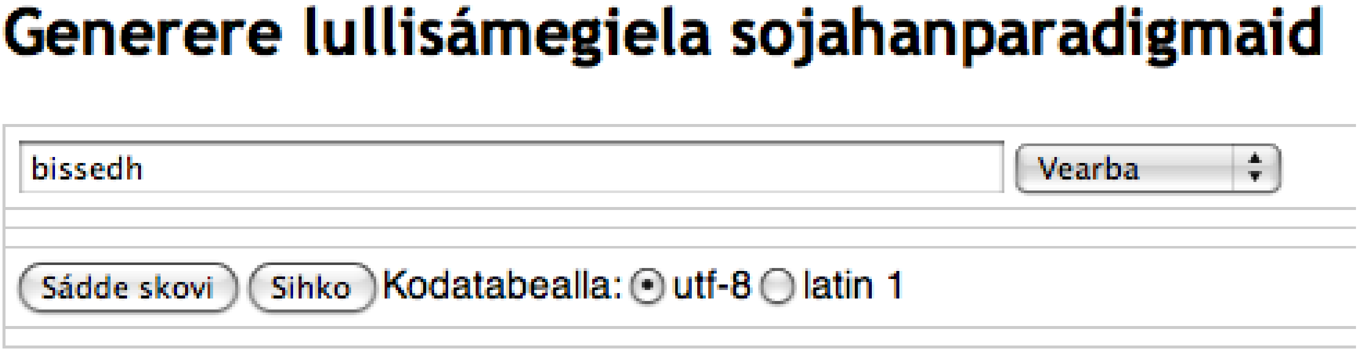

gussa+N+Sg+Acc
gussa^WeG
gussa^WeG
gusa



```
gusa
gusa    gussa+N+Sg+Acc
gusa    gussa+N+Sg+Gen

girjji
girjji  girji+N+Sg+Acc
girjji  girji+N+Sg+Gen

girjái
girjái  girji+N+Sg+Ill
girjái  girjái+A+Sg+Ill
girjái  girjái+A+Sg+Nom
```

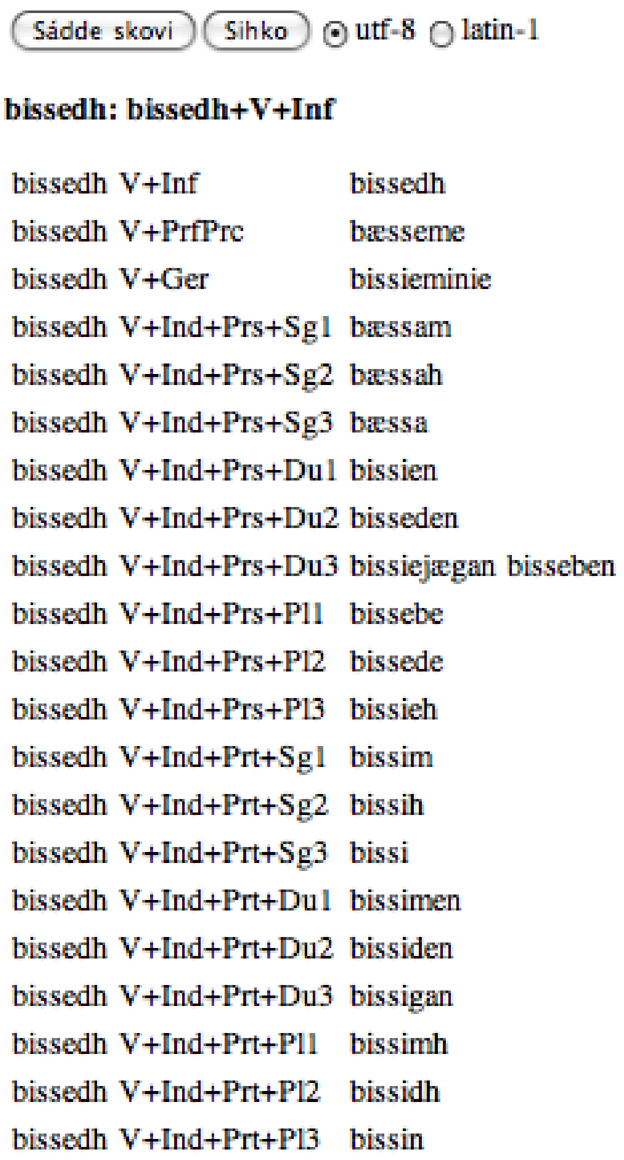


Generere lullisámegiela sojahanparadigmaid

bissedh

Vearba

Kodatabealla: ● utf-8 ○ latin 1



● utf-8 ○ latin-1

bissedh: bissedh+V+Inf

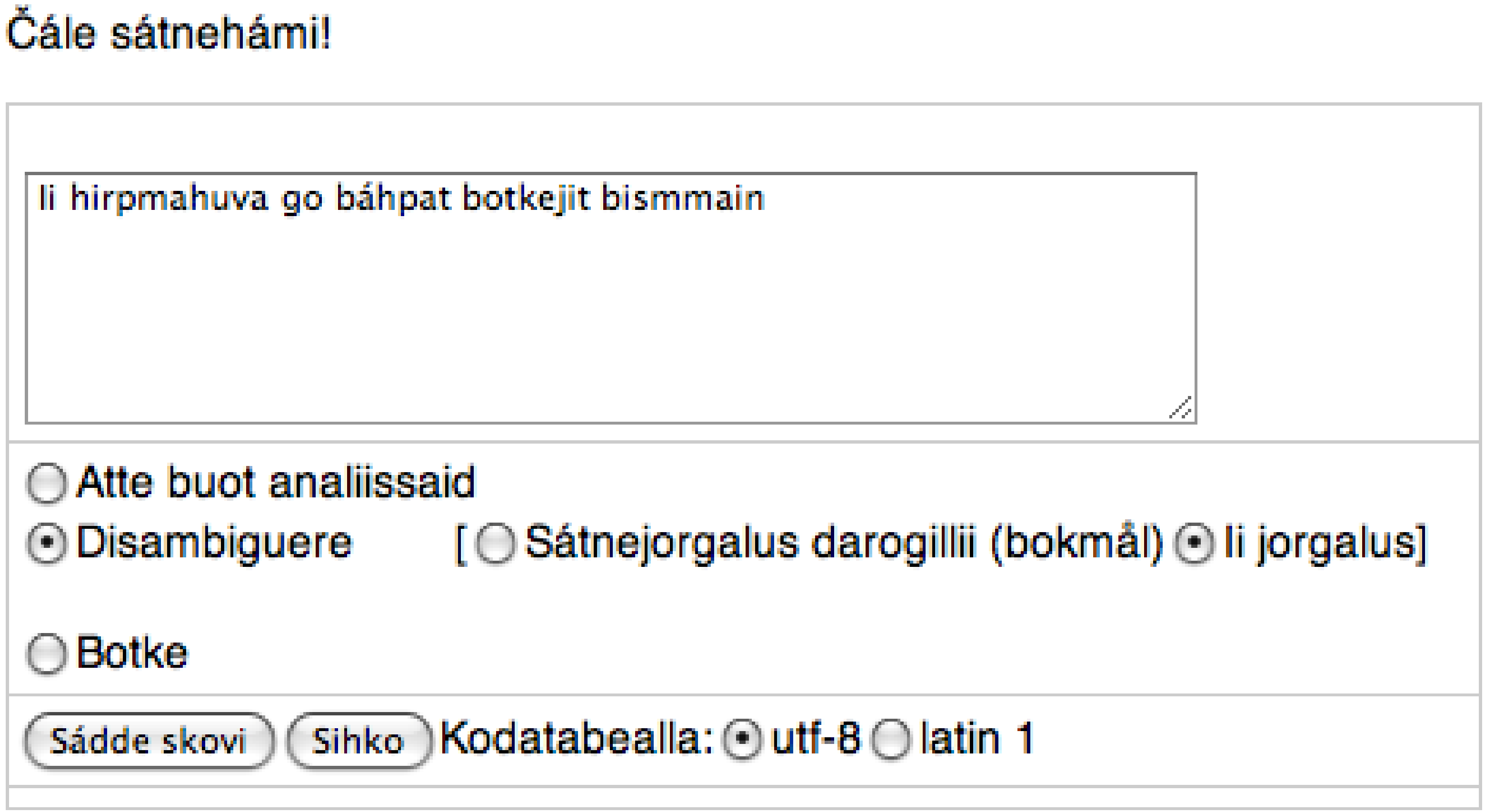
```
bissedh V+Inf            bissedh
bissedh V+PrfPrc         bæsseme
bissedh V+Ger            bissieminie
bissedh V+Ind+Prs+Sg1    bæssam
bissedh V+Ind+Prs+Sg2    bæssah
bissedh V+Ind+Prs+Sg3    bæssa
bissedh V+Ind+Prs+Du1    bissien
bissedh V+Ind+Prs+Du2    bisseden
bissedh V+Ind+Prs+Du3    bissiejægan bisseben
bissedh V+Ind+Prs+Pl1    bissebe
bissedh V+Ind+Prs+Pl2    bissede
bissedh V+Ind+Prs+Pl3    bissieh
bissedh V+Ind+Prt+Sg1    bissim
bissedh V+Ind+Prt+Sg2    bissih
bissedh V+Ind+Prt+Sg3    bissi
bissedh V+Ind+Prt+Du1    bissimen
bissedh V+Ind+Prt+Du2    bissiden
bissedh V+Ind+Prt+Du3    bissigan
bissedh V+Ind+Prt+Pl1    bissimh
bissedh V+Ind+Prt+Pl2    bissidh
bissedh V+Ind+Prt+Pl3    bissin
```

Morphological disambiguators

- Ambiguous words become clear in context
- → Constraint grammar
- → Manually written ruleset

Syntactic analysers

- Adding grammatical function and dependency



Čále sátnehámi!

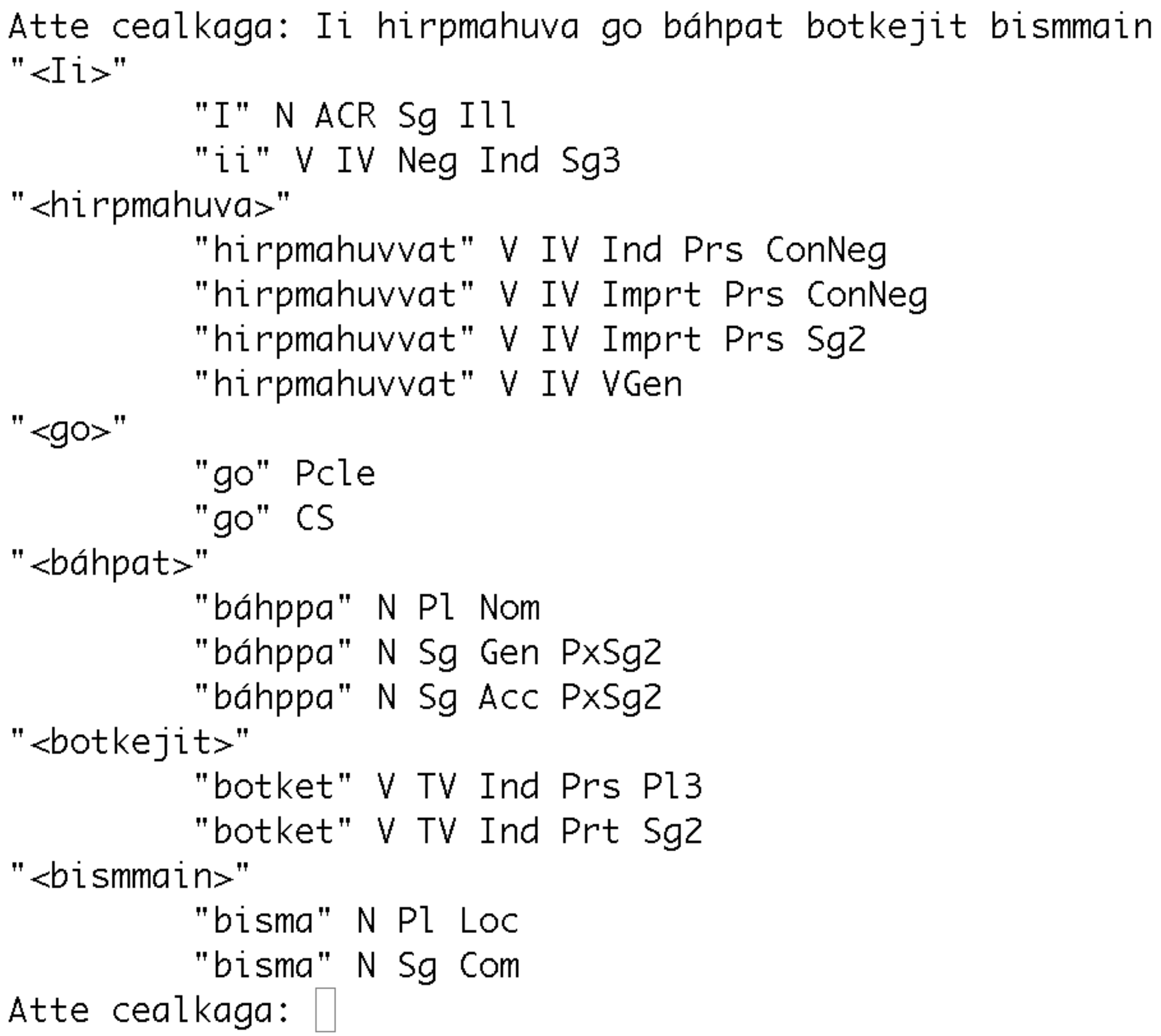
Ii hirpmahuva go báhpat botkejit bismmain

○ Atte buot analiissaid

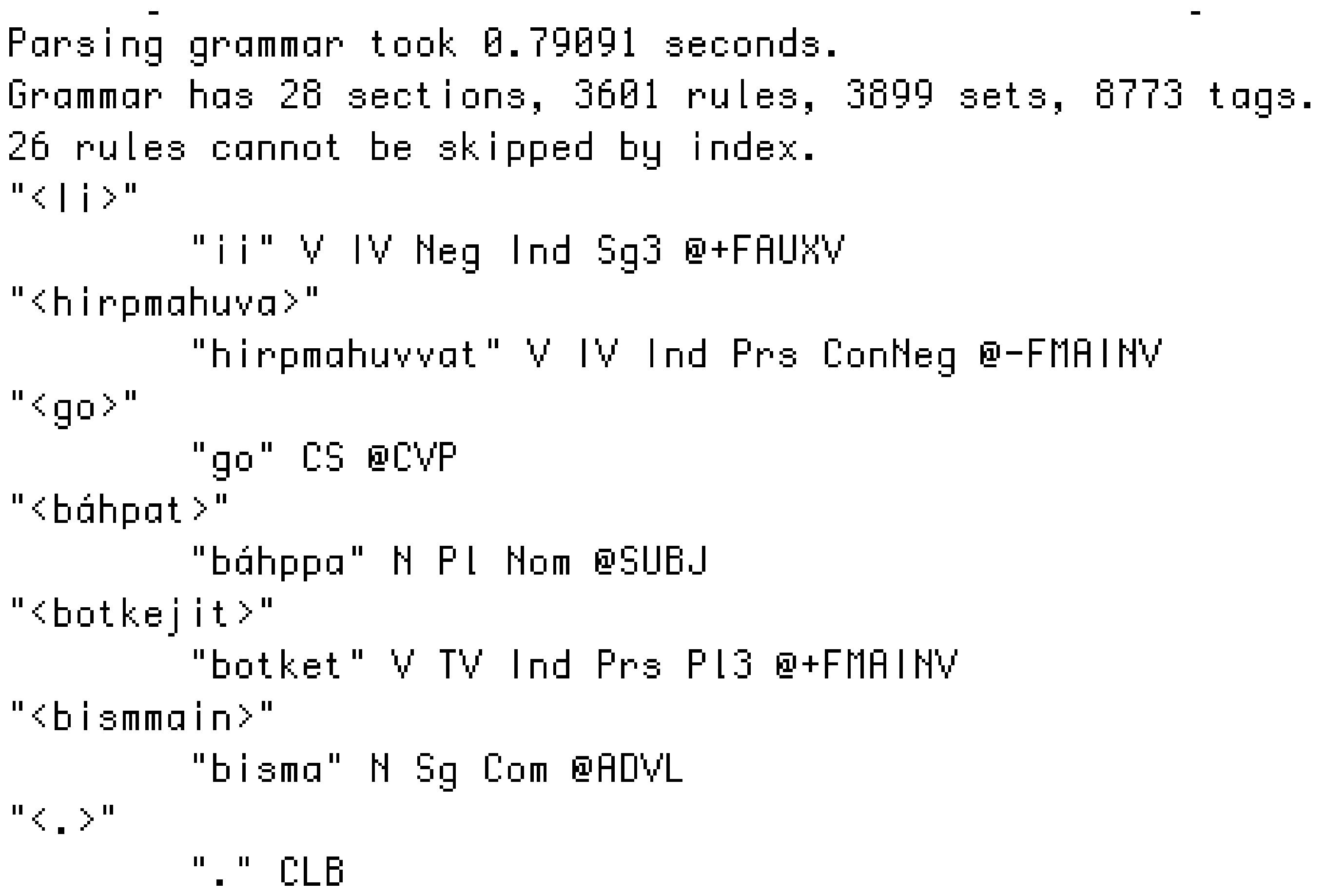
● Disambiguere [○ Sátnejorgalus darogillii (bokmål) ● Ii jorgalus]

○ Botke

Kodatabealla: ● utf-8 ○ latin 1



```
Atte cealkaga: Ii hirpmahuva go báhpat botkejit bismmain
"<Ii>"
        "I" N ACR Sg Ill
        "ii" V IV Neg Ind Sg3
"<hirpmahuva>"
        "hirpmahuvvat" V IV Ind Prs ConNeg
        "hirpmahuvvat" V IV Imprt Prs ConNeg
        "hirpmahuvvat" V IV Imprt Prs Sg2
        "hirpmahuvvat" V IV VGen
"<go>"
        "go" Pcle
        "go" CS
"<báhpat>"
        "báhppa" N Pl Nom
        "báhppa" N Sg Gen PxSg2
        "báhppa" N Sg Acc PxSg2
"<botkejit>"
        "botket" V TV Ind Prs Pl3
        "botket" V TV Ind Prt Sg2
"<bismmain>"
        "bisma" N Pl Loc
        "bisma" N Sg Com
Atte cealkaga:
```



```
Parsing grammar took 0.79091 seconds.
Grammar has 28 sections, 3601 rules, 3899 sets, 8773 tags.
26 rules cannot be skipped by index.
"<li>"
        "ii" V IV Neg Ind Sg3 @+FAUXV
"<hirpmahuva>"
        "hirpmahuvvat" V IV Ind Prs ConNeg @-FMAINV
"<go>"
        "go" CS @CVP
"<báhpat>"
        "báhppa" N Pl Nom @SUBJ
"<botkejit>"
        "botket" V TV Ind Prs Pl3 @+FMAINV
"<bismmain>"
        "bisma" N Sg Com @ADVL
"<.>"
        "." CLB
```

Why do we use just these methods, and not other?

Let us have a quick look at the alternatives

The alternatives

- Morphology
  - fullform lists
  - shallow parsing (part of speech only)
- Disambiguation and syntax
  - "deeper" syntactic approaches: LFG, HPSG
  - "more shallow" approaches: statistical disambiguators

Morphology

For languages with...

- less morphology, *morphfeature:wordform* pairs are ok
- extensive but concatenative morphology, simple automata are ok
- extensive **and** non-concatenative morphology, we find cascading or two-level transducers the best option

POS-only information is good for some applications, we want to know that *gusa* → *gussa*

The real reason why we do it our way:

A transducer model of a grammar is a generative grammar of the language in question

By using finite state transducers rather than wordlist approachess, we are as linguists able to test our grammatical hypotheses in full scale, rather than on a couple of examples

→ **Here we have a substantial motivation for spending years on making a program with no commercial potential**

When should a language get such a transducer?

- Linguistically speaking, *always*
- As part of a revitalisation project: Perhaps not the first thing to do
- (1st priority for lingvists is grammar - dictionary - text collection)

- For university lingustics, languages with few speakers are as interesting as languages with many speakers
- Even more so: Languages where you may be a pioneer may be more attractive
- → **Anyone interested in Inari Sámi consonant gradation should make a comprehensive finite state transducer**

Disambiguation and syntax

Our philosophy:

1. What we do shall be linguistically interesting
    - → not statistical disambiguating
    - they work (96 % accuracy), but do not tell us about the grammar

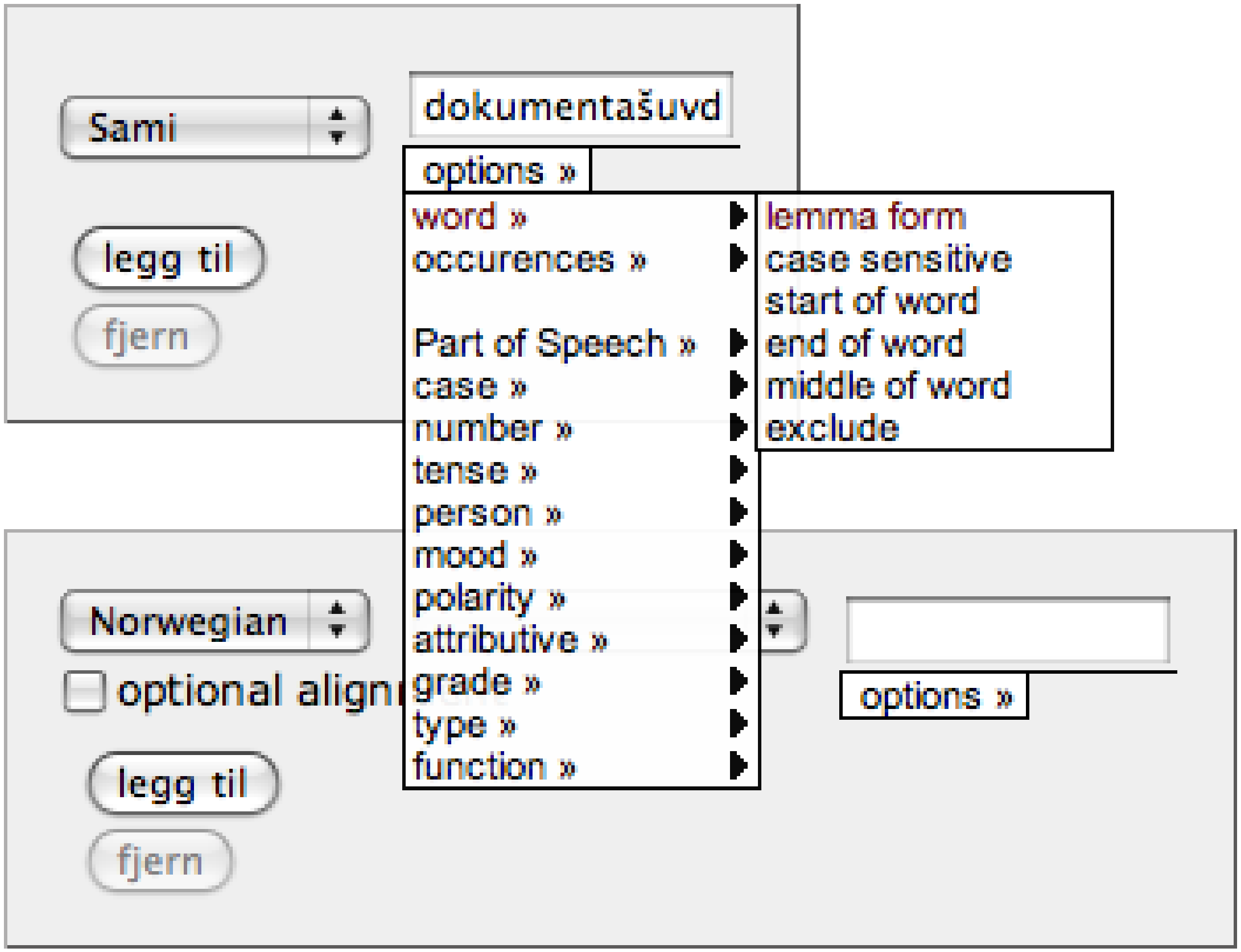
2. What we do must work (be robust)
    - → hence not syntactic models such as *LFG, HPSG*
    - They give very good analyses, but only in 60 % of the cases
    - They often give thousands of analysis for the same sentence

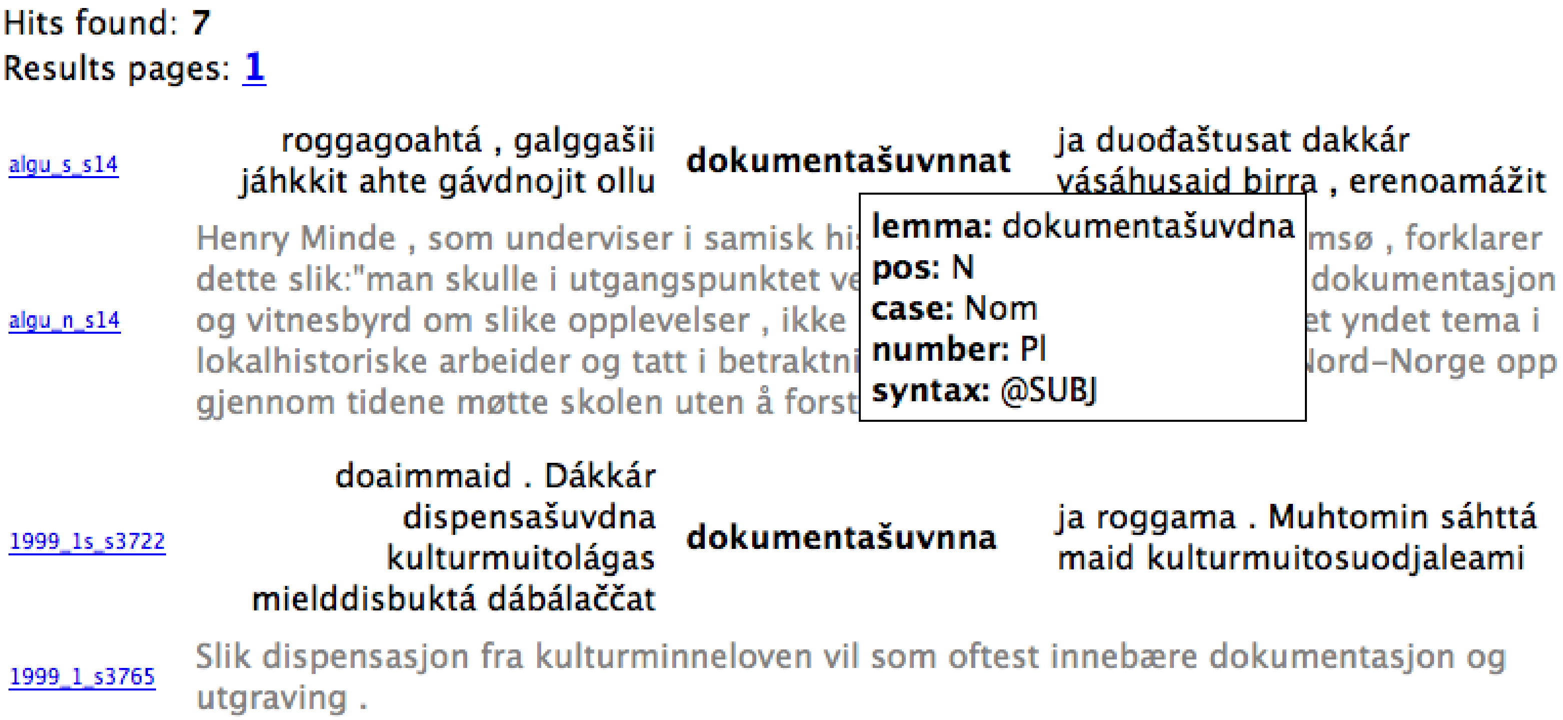
Constraint grammar is our choice

- → It has a higher accuracy (97-99 % reported)
- → We add grammatical function and dependency, and approach the level of deep parsers in information richness

Thus, we have a robust syntactic analyser

- basis for end-user applications such as grammar checking and machine translation
- also relevant to lexicographers, terminologists, ...


Sami
dokumentašuvd
options »
legg til
fjern
word »
occurences »
Part of Speech »
case »
number »
tense »
person »
mood »
polarity »
attributive »
grade »
type »
function »
lemma form
case sensitive
start of word
end of word
middle of word
exclude
Norwegian
optional align
legg til
fjern
options »



Hits found: **7**
Results pages: 1

algu_s_s14
roggagoahtá , galggašii jáhkkit ahte gávdnojit ollu **dokumentašuvnnat** ja duođaštusat dakkár vásáhusaid birra , erenoamážit

lemma: dokumentašuvdna
pos: N
case: Nom
number: Pl
syntax: @SUBJ

algu_n_s14
Henry Minde , som underviser i samisk hi msø , forklarer
dette slik:"man skulle i utgangspunktet ve dokumentasjon
og vitnesbyrd om slike opplevelser , ikke et yndet tema i
lokalhistoriske arbeider og tatt i betraktni Nord–Norge opp
gjennom tidene møtte skolen uten å forst

1999_1s_s3722
doaimmaid . Dákkár dispensašuvdna kulturmuitolágas mielddisbuktá dábálaččat **dokumentašuvnna** ja roggama . Muhtomin sáhttá maid kulturmuitosuodjaleami

1999_1_s3765
Slik dispensasjon fra kulturminneloven vil som oftest innebære dokumentasjon og utgraving .

Lene

Our corpus

Existing dialect materials

- Written dialect texts
- Dialect recordings
- A small part is transcripted
- Different methods for transcription

A corpus of spoken Sámi?

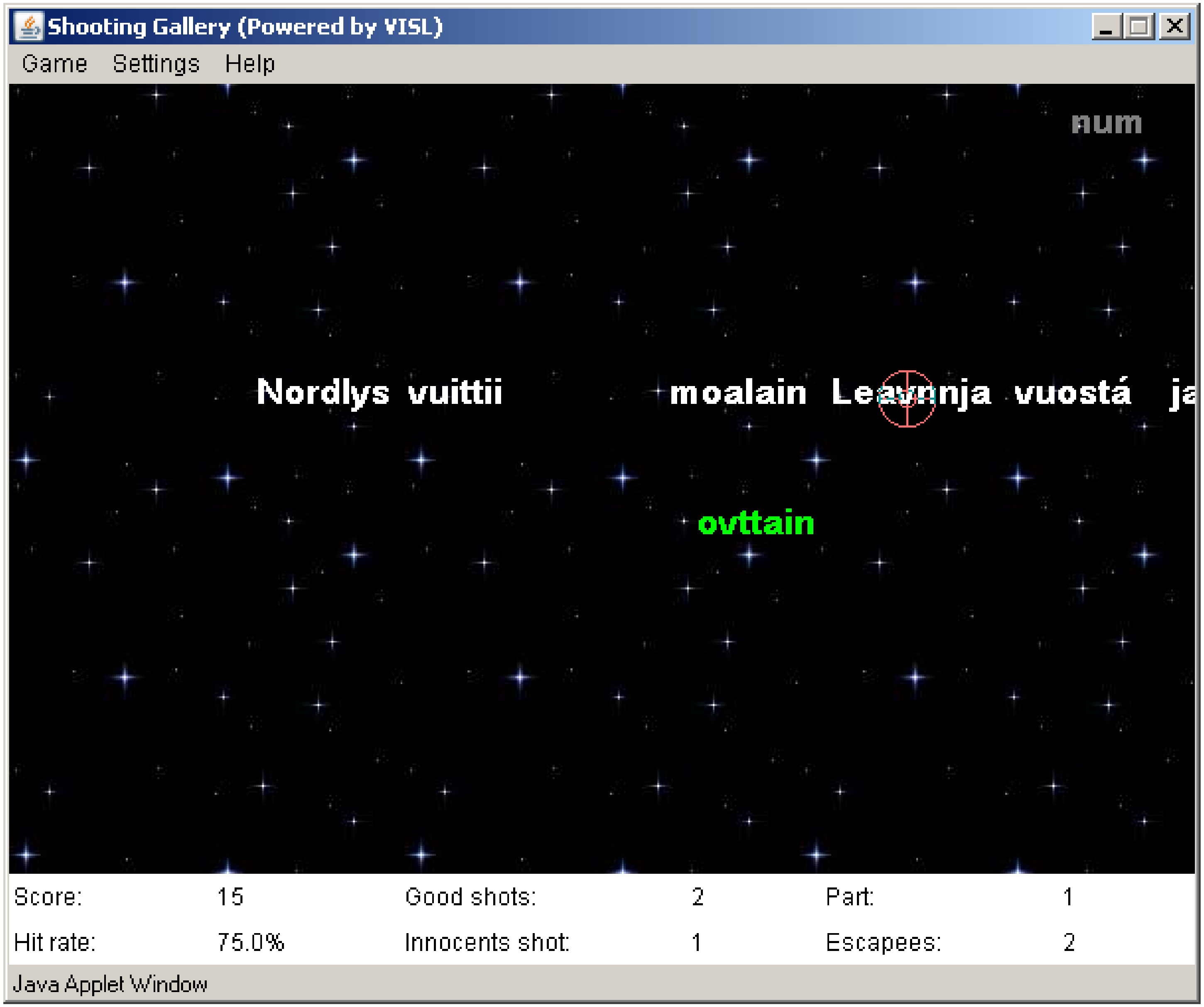
- Collect the transcriptions, transcribe more
- Automatic conversion into standard orthography → grammatical analysis
- Parallell corpus: transcription / standard orthography
- Recordings available

Pedagogical programs - based upon language technology

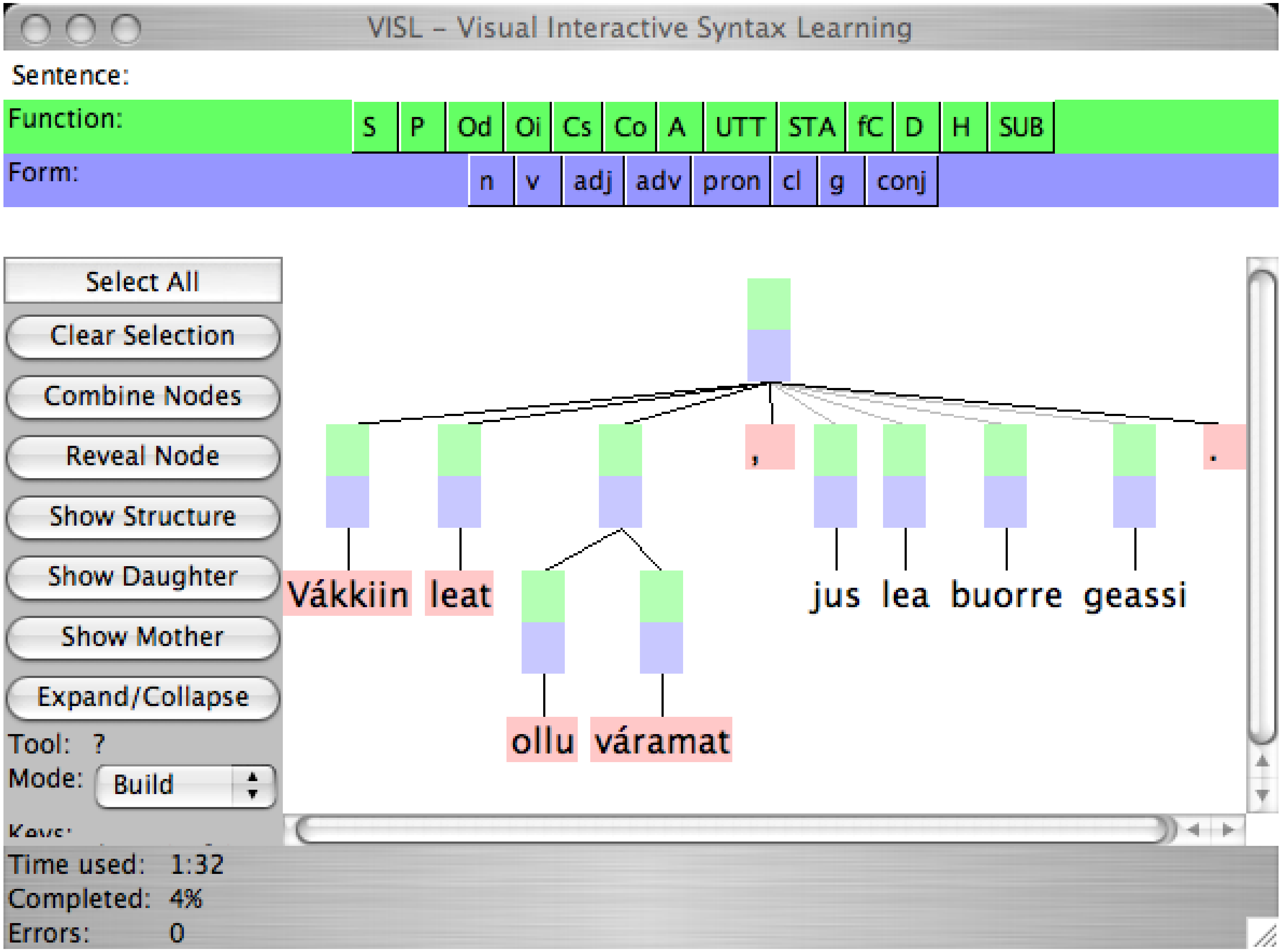
Two goals

1. interactive grammar games (the technology by VISL - University of Southern Denmark)
2. make programs for interactive grammar and communicative exercises in Sámi

(The project is funded by the Faculty of Humanities at UiT and the Sámi parlament in Norway)

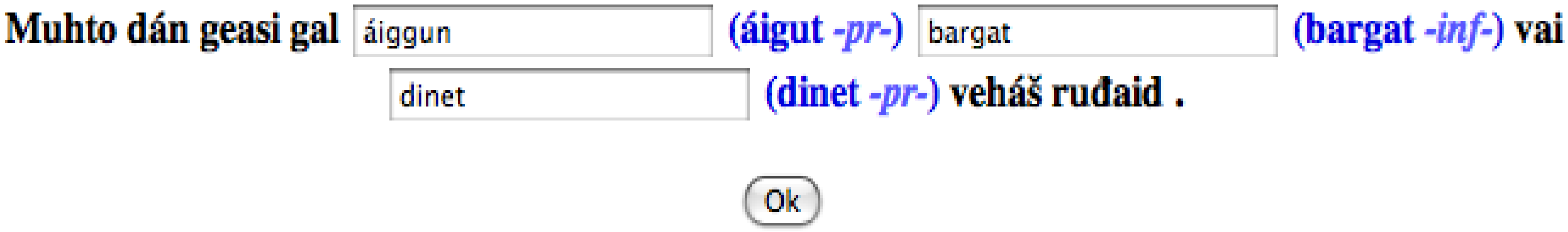


Score: 15
Good shots: 2
Part: 1
Hit rate: 75.0%
Innocents shot: 1
Escapees: 2

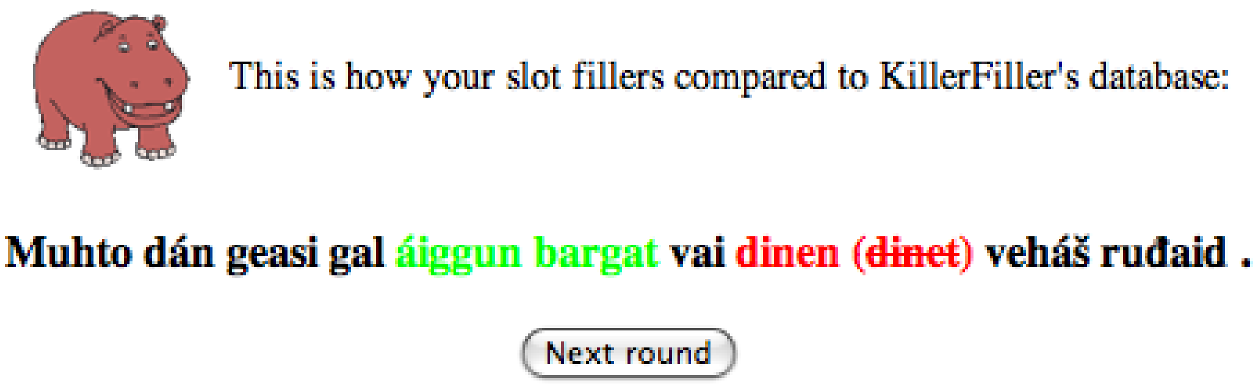


Tool: ?
Mode: Build

Time used: 1:32
Completed: 4%
Errors: 0



Muhto dán geasi gal áiggun (áigut *-pr-*) bargat (bargat *-inf-*) vai dinet (dinet *-pr-*) veháš ruđaid .

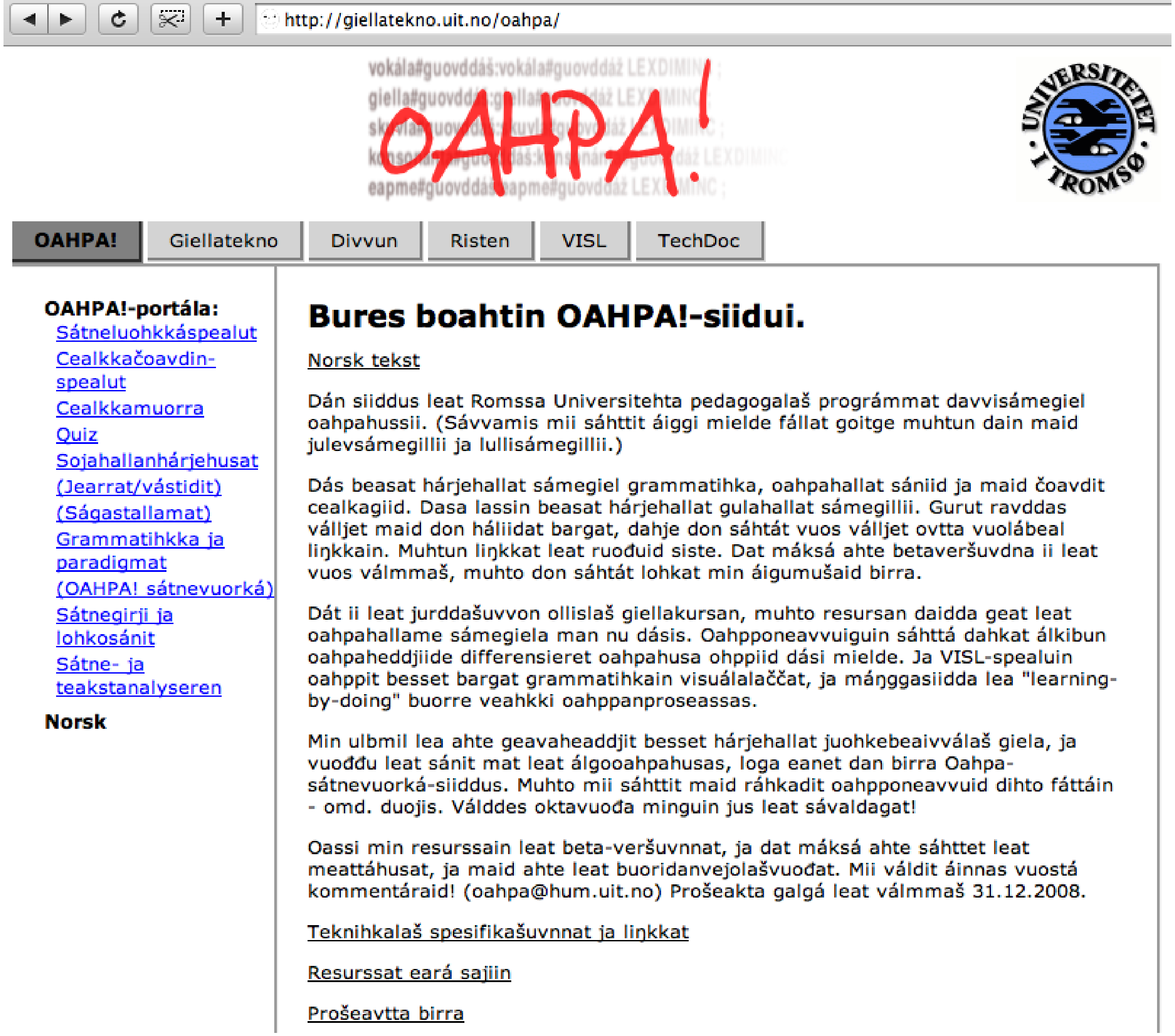


This is how your slot fillers compared to KillerFiller's database:

Muhto dán geasi gal áiggun bargat vai dinen (~~dinet~~) veháš ruđaid .

The dialogues and drills are based upon our lexica and analysers:

- *Mas don balat? (What are you afraid of?)*
  - May accept all answers containing locative, both singular and plural, also negative
- *Jugat go gáfe? (Do you drink coffee?)*
  - May accept all answers containing "juhkat" 2Sg presens, both indicative and conditional, both affirming and negative



Bures boahtin OAHPA!-siidui.

Norsk tekst

Dán siiddus leat Romssa Universitehta pedagogalaš prográmmat davvisámegiel oahpahussii. (Sávvamis mii sáhttit áiggi mielde fállat goitge muhtun dain maid julevsámegillii ja lullisámegillii.)

Dás beasat hárjehallat sámegiel grammatihka, oahpahallat sániid ja maid čoavdit cealkagiid. Dasa lassin beasat hárjehallat gulahallat sámegillii. Gurut ravddas válljet maid don háliidat bargat, dahje don sáhtát vuos válljet ovtta vuolábeal liŋkkain. Muhtun liŋkkat leat ruođuid siste. Dat máksá ahte betaveršuvdna ii leat vuos válmmaš, muhto don sáhtát lohkat min áigumušaid birra.

Dát ii leat jurddašuvvon ollislaš giellakursan, muhto resursan daidda geat leat oahpahallame sámegiela man nu dásis. Oahpponeavvuiguin sáhttá dahkat álkibun oahpaheddjiide differensieret oahpahusa ohppiid dási mielde. Ja VISL-spealuin oahppit besset bargat grammatihkain visuálalaččat, ja máŋggasiidda lea "learning-by-doing" buorre veahkki oahppanproseassas.

Min ulbmil lea ahte geavaheaddjit besset hárjehallat juohkebeaivválaš giela, ja vuođđu leat sánit mat leat álgooahpahusas, loga eanet dan birra Oahpa-sátnevuorká-siiddus. Muhto mii sáhttit maid ráhkadit oahpponeavvuid dihto fáttáin - omd. duojis. Válddes oktavuođa minguin jus leat sávaldagat!

Oassi min resurssain leat beta-veršuvnnat, ja dat máksá ahte sáhttet leat meattáhusat, ja maid ahte leat buoridanvejolašvuođat. Mii váldit áinnas vuostá kommentáraid! (oahpa@hum.uit.no) Prošeakta galgá leat válmmaš 31.12.2008.

Teknihkalaš spesifikašuvnnat ja liŋkkat

Resurssat eará sajiin

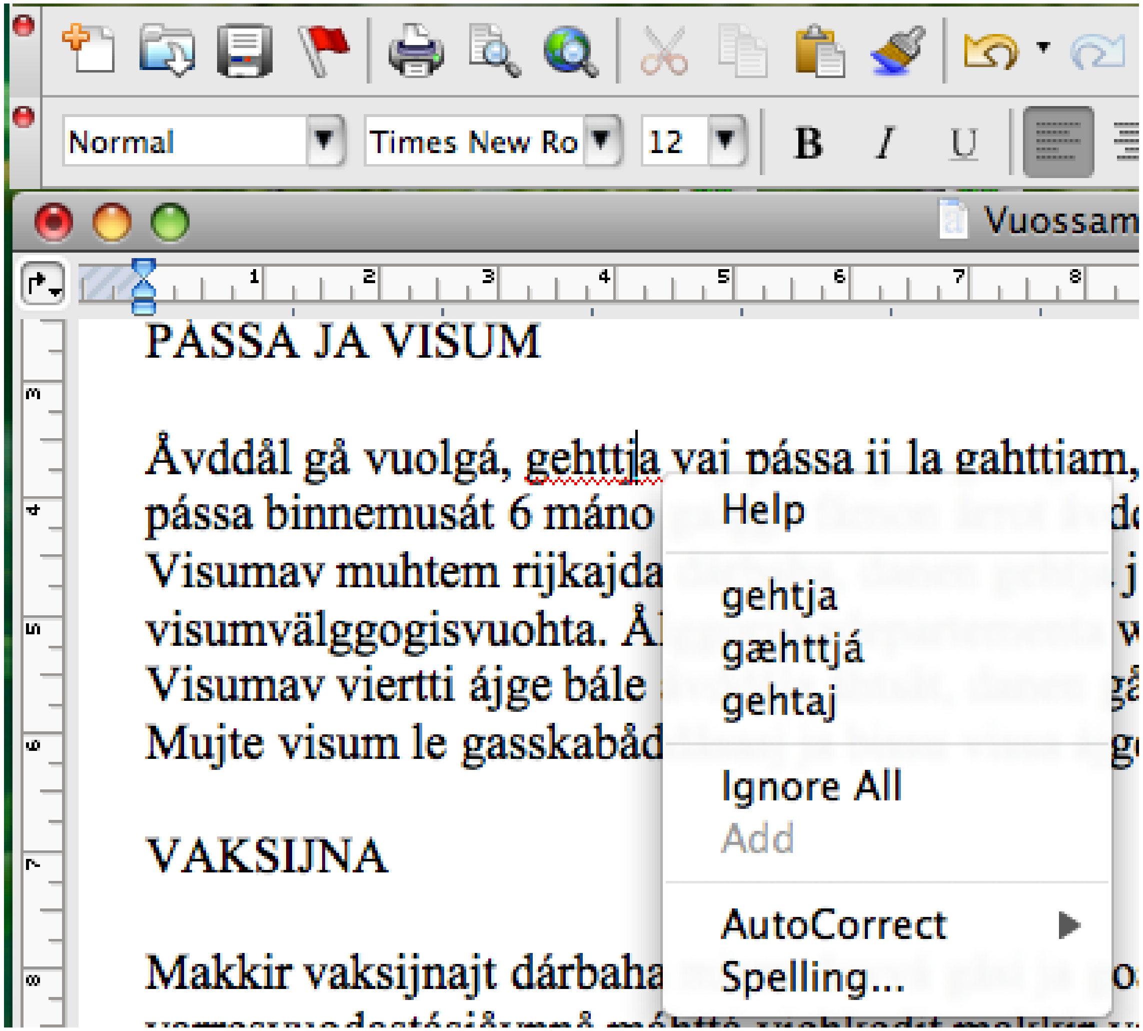
Prošeavtta birra

Sjur

Proofing tools

- Spell checker
- Hyphenator
- Possible in the future:
  - inflecting thesaurus
  - grammar checker

Spell checker



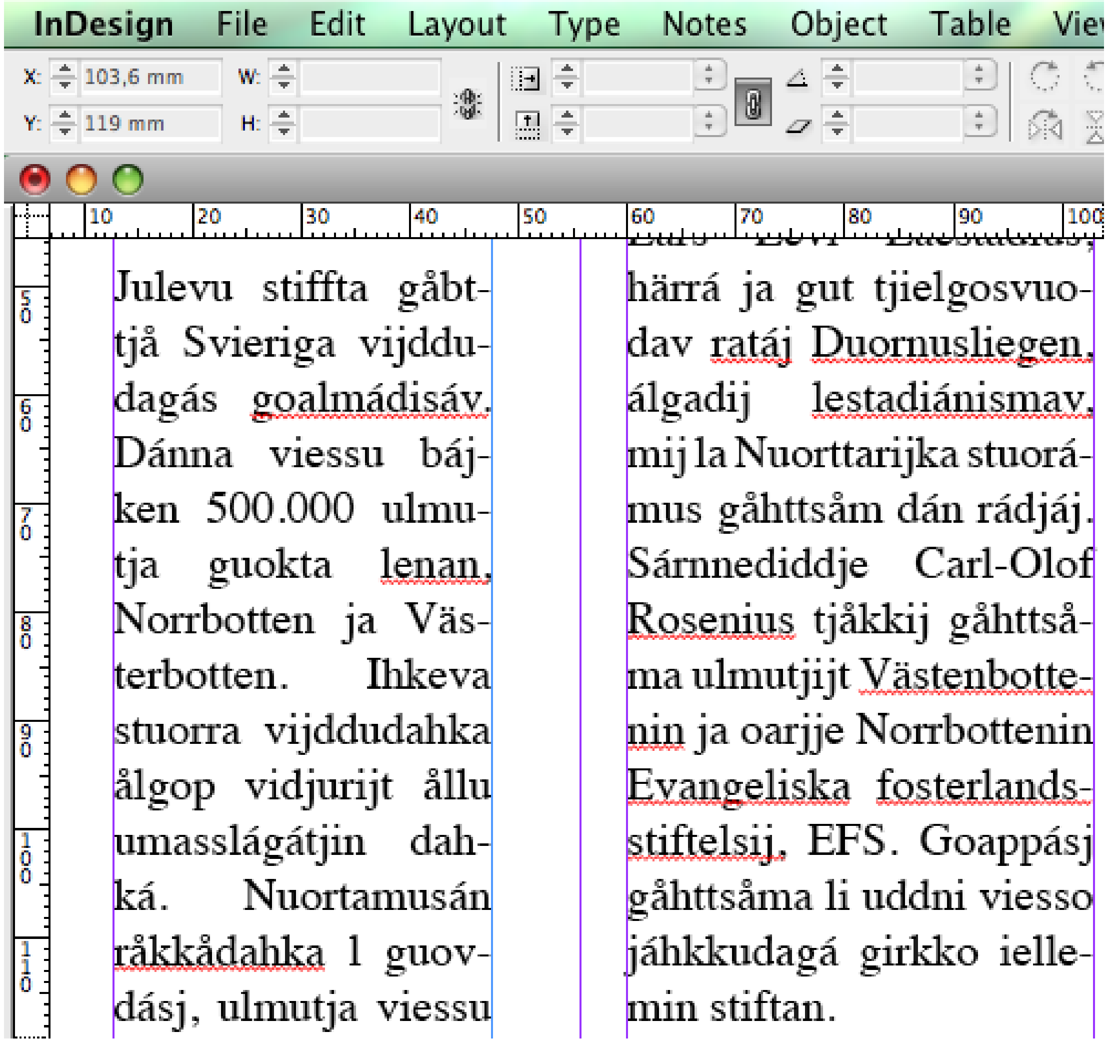
PASSA JA VISUM

Åvddål gå vuolgá, gehttja vai pássa ii la gahttiam,
pássa binnemusát 6 máno
Visumav muhtem rijkajda
visumvälggogisvuohta. Å
Visumav viertti ájge bále
Mujte visum le gasskabåd

VAKSIJNA

Makkir vaksijnajt dárbaha

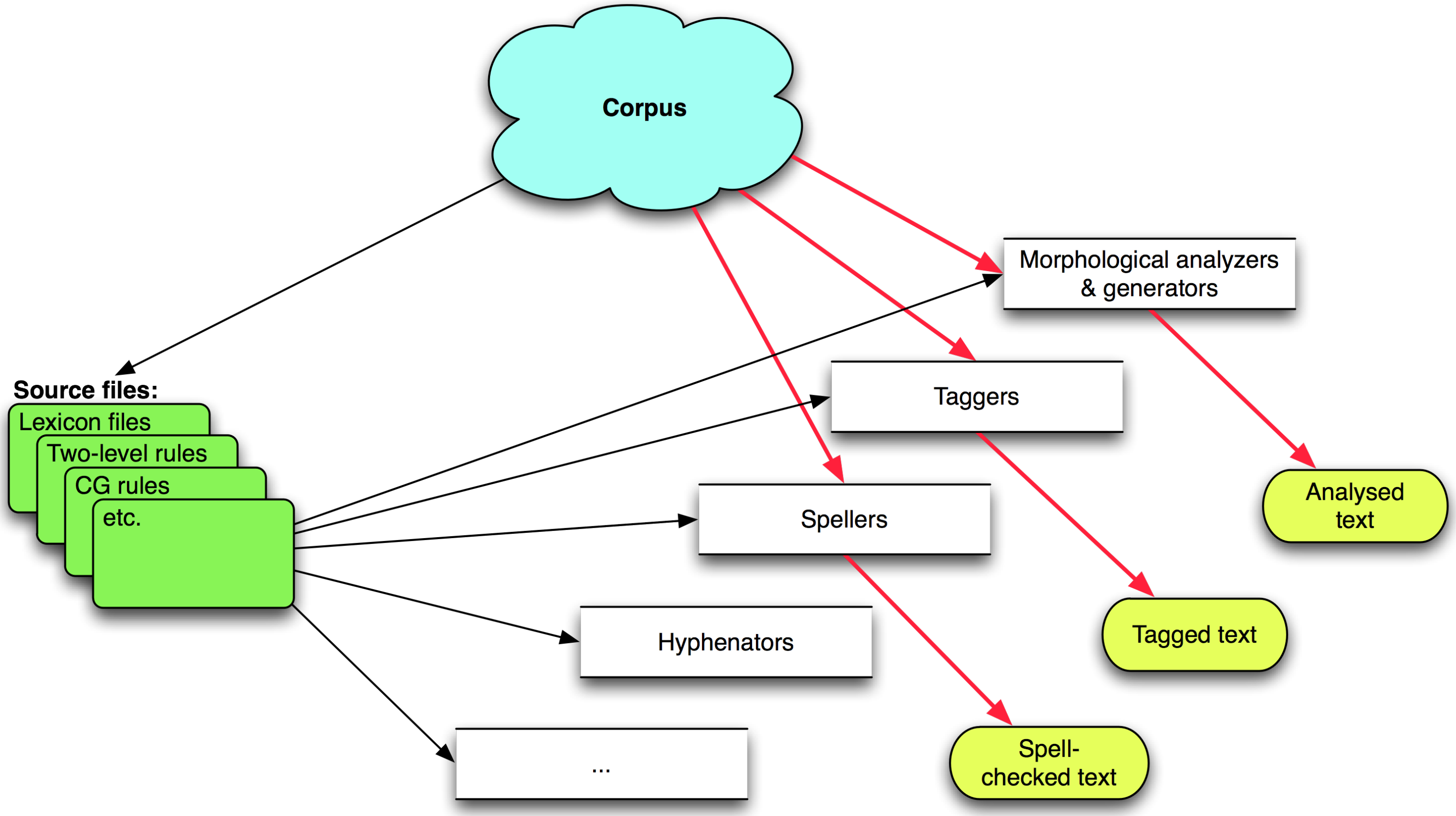
Automatic hyphenation



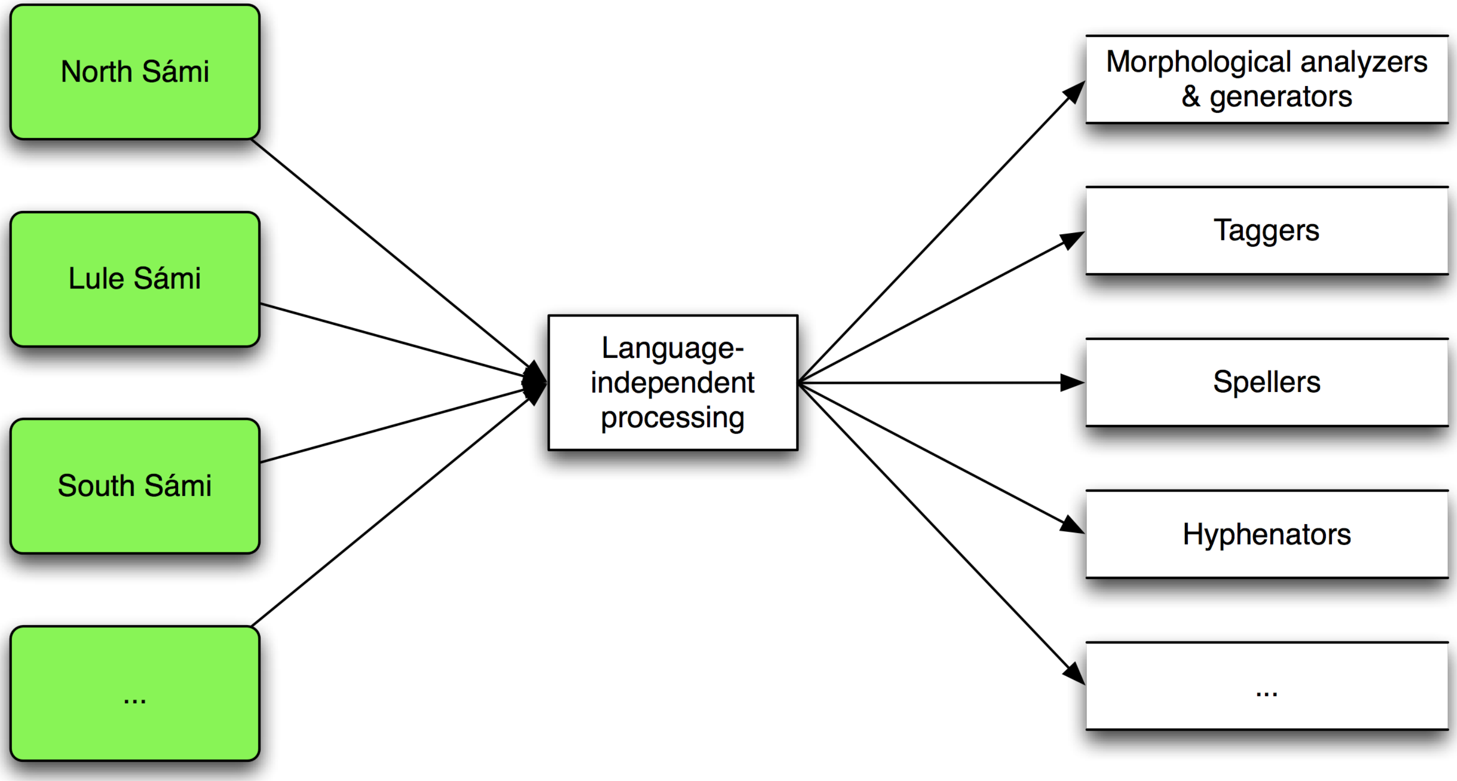
Julevu stiffta gåbt-
tjå Svieriga vijddu-
dagás goalmádisáv.
Dánna viessu báj-
ken 500.000 ulmu-
tja guokta lenan,
Norrbotten ja Väs-
terbotten. Ihkeva
stuorra vijddudahka
ålgop vidjurijt ållu
umasslágátjin dah-
ká. Nuortamusán
råkkådahka 1 guov-
dásj, ulmutja viessu

Lars Levi Laestadius,
härrá ja gut tjielgosvuo-
dav ratáj Duornusliegen,
álgadij lestadiánismav,
mij la Nuorttarijka stuorá-
mus gåhttsåm dán rádjáj.
Sárnnediddje Carl-Olof
Rosenius tjåkkij gåhttså-
ma ulmutjijt Västenbotte-
nin ja oarjje Norrbottenin
Evangeliska fosterlands-
stiftelsij, EFS. Goappásj
gåhttsåma li uddni viesso
jáhkkudagá girkko ielle-
min stiftan.

What ties it all together?



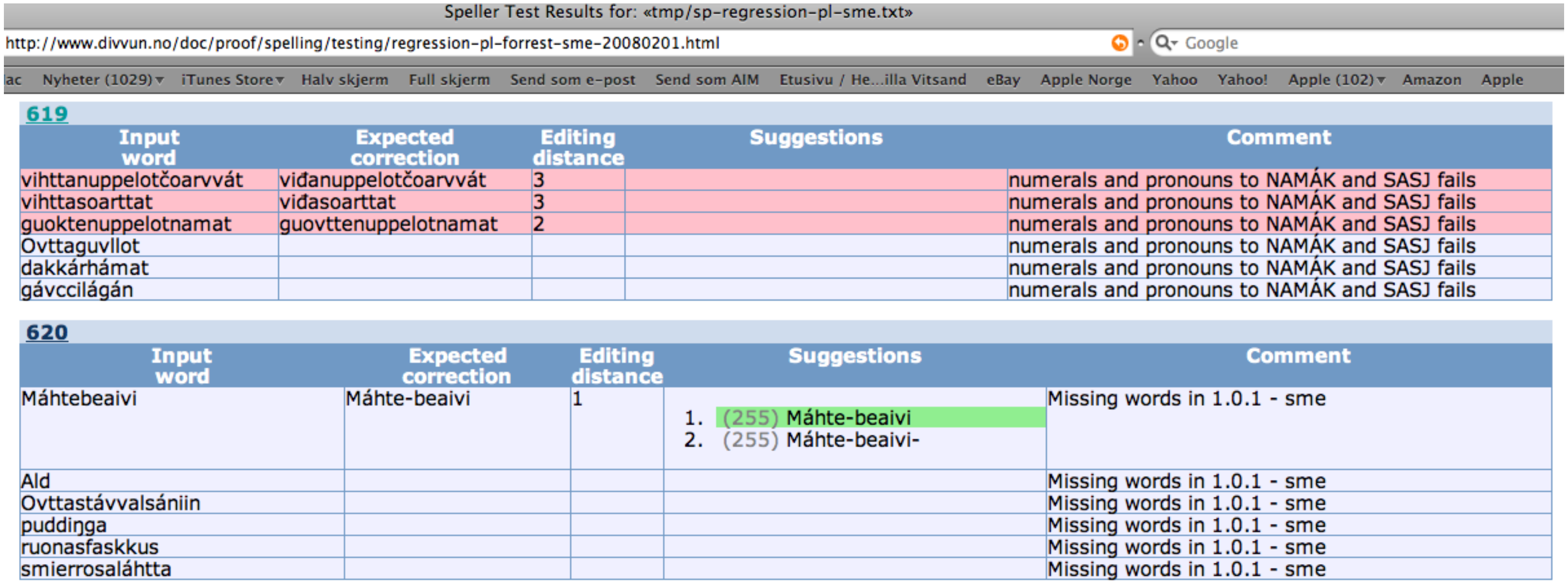
Development Infrastructure


North Sámi
Lule Sámi
South Sámi
...
Language-independent processing
Morphological analyzers & generators
Taggers
Spellers
Hyphenators
...

Testing Infrastructure

- Two-level rule test pairs
- Lexical coverage / corpus analysis
- Proofing tools testing:
  - Gold standard testing (precision/recall)
  - Regression testing
  - Typos testing

Regression testing



Speller Test Results for: «tmp/sp-regression-pl-sme.txt»

http://www.divvun.no/doc/proof/spelling/testing/regression-pl-forrest-sme-20080201.html

619

Input word	Expected correction	Editing distance	Suggestions	Comment
vihttanuppelotčoarvvát	viđanuppelotčoarvvát	3		numerals and pronouns to NAMÁK and SASJ fails
vihttasoarttat	viđasoarttat	3		numerals and pronouns to NAMÁK and SASJ fails
guoktenuppelotnamat	guovttenuppelotnamat	2		numerals and pronouns to NAMÁK and SASJ fails
Ovttaguvllot				numerals and pronouns to NAMÁK and SASJ fails
dakkárhámat				numerals and pronouns to NAMÁK and SASJ fails
gávccilágán				numerals and pronouns to NAMÁK and SASJ fails

620

Input word	Expected correction	Editing distance	Suggestions	Comment
Máhtebeaivi	Máhte-beaivi	1	1. (255) Máhte-beaivi 2. (255) Máhte-beaivi-	Missing words in 1.0.1 - sme
Ald				Missing words in 1.0.1 - sme
Ovttastávvalsániin				Missing words in 1.0.1 - sme
puddiŋga				Missing words in 1.0.1 - sme
ruonasfaskkus				Missing words in 1.0.1 - sme
smierrosaláhtta				Missing words in 1.0.1 - sme

Portability

Goal: Port solutions for Northern Sámi to other languages

- Large costs go into setting up infrastructure.
- Commercial companies naturally keep this infrastructure to themselves, as this is part of their competitive advantage
- In Tromsø, we publish our infrastructure as part of an open-source *how-to* for language technology projects.

Conclusion: Language technology solutions are

- a *sine qua non* for minority languages needing a written language
- necessary tools for reference work.
- Linguists, programmers and language activists should co-operate on making the necessary tools for supporting use of the literary language