UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

Spoken language corpora

Janne Bondi Johannessen

AAU, March 2014

Spoken corpora at the Text Laboratory, UiO

- Big Brother corpus
  - 2001. ca. 550 000 words
- Nordic dialect corpus
  - 2012. Ca. 2 mill. words
- NoTa-Oslo
  - 2006. Ca. 900 000 words
- Ruija corpus
  - 2006. Ca. 400 000 words.
- TAUS
  - 1975 / 2006Old Oslo. Ca. 250 000 words.
- Doctor patient corpus
  - 2014

Spoken language corpora

- Purposes of a spoken language corpus
  - Study aspects of language as with written corpora
  - In addition, studies of:
    - Discourse
    - Phonology
    - Dialect
    - Diachronic change
    - Agelect
    - ...

How make spoken data into a corpus?

- Digitise sound/video
- Transcribe speech
  - Choose transcription
    - Phonetic
    - Orthographic
- Time code transcription for alignment with sound file
- Annotate
  - Extra-linguistic features (laughter, coughing)
  - Semi-linguistc features (meaningful sounds)
  - Gestures
  - Grammatically
    - POS tagging
    - Syntactic parsing
    - Other

Different situation and informants - different speech types

- The Big Brother Corpus
- Pros
  - Lots of spontaneous speech data
  - Lots of dialogue and polylogue
  - Lots of emotional speech in different dialogue situations
    - Conflict, argument, love, irritation etc.
- Cons
  - Not a dialect corpus
  - No representativity w.r.t. age, social class, education etc.
  - Not ”controlled” recording situations
  - Small number of informants

- Nordic Dialect Corpus
- Pros
  - Lots of spontaneous speech data
  - Lots of dialogue, no polylogue
  - No emotional speech
- Cons
  - Is a dialect corpus
  - Partly representative
  - Controlled recording situation
  - Small AND large number of informants

Big Brother vs Nordic Dialect Corpus

Trouble viewing video?
context±
Offset
Left −1
Right 1
- Start +
- Stop +
>>

Trouble viewing video?
context±
Offset
Left −1
Right 1
- Start +
- Stop +
>>

Other dialect corpora?

We know of no comparable resource for any language

- *Sounds familiar? Accents and Dialects of the UK*
  - No grammatical search options
  - No results handling
- *The British National Corpus*
  - No audio
  - Orthographic transcription
  - Unreliable dialect categories
- *The DynaSand dialect database*
  - Few spontaneous utterances
- *The Spoken Dutch Corpus*
  - Not web-based
  - Orthographic transcriptions
  - Not dialect data
- The Scottish Corpus of Text and Speech
  - Not a dialect corpus
  - No searchable linguistic features
- Others under development:
  - Corpus of Estonian Dialects
  - Spoken Japanese Dialect Corpus
- Paul Thompson at the University of Reading: Posting at Corpora List 30 Nov. 2008 about linked audio or video files with transcripts: 15 answers, of which only one on dialects: ours

Comparison: NoTa-Oslo, Talesøk, BySoc, BNC, SCOTS

	NoTa	Talesøk	GSLC	BySoc	BNC	SCOTS
Transcription linked to audio	Yes	Yes	No	No	No	Yes
Transcription linked to video	Yes	No	No	No	No	Yes
User-friendly search without regular expressions	Yes	Yes	No	No	No	No
Possible to limit informant selection	Yes	Yes	Yes	Yes	Yes	Yes
Overlaps/ turntaking annotated	Yes	Yes	Yes	Yes	No	Yes
Transcription as standard orthography (or slightly modified)	Yes	Yes	Yes	Yes	Yes	Yes
POS tagged	Yes	No	No	No	Yes	No
POS tags can be used as the only search expressions	Yes	–	–	–	No	–

Credits

- Collaboration between the research network ScanDiaSyn and Nordic Centre of Excellence in Microcomparative Syntax (NORMS)
- Some of the corpus is financed by national research councils in the individual countries
- The technical development has been financed by the University of Oslo, the Norwegian Research Council, and the Nordic research funds NOS-HS and NordForsk

Why the Nordic Dialect Corpus was developed

- Initiated by members of Nordic Centre of Excellence in Microcomparative Syntax and the ScanDiaSyn network
- Overarching goal: to study the dialects of the North-Germanic dialect continuum
  - The Nordic languages are closely related and have some mutually intelligibility
  - Studying the dialects within each national language is misguided from a theoretical and principled point of view
  - Difficult for each researcher to get hold of relevant data on their own in such a large area.
  - Many different kinds of data needed for syntax research

Corpus features

- Linguistic contents
  - Dialects from five closely related languages
- Annotation
  - POS tagging and two types of transcription
- Search interface
  - Advanced possibility to combine an array of search criteria and results presentation in an intuitive interface
- Many search variables
  - Linguistics-based, informant-based, time-based
- Multimedia display
  - Linking of sound and video to transcriptions
- Display of informant details
  - Number of words and other informant information
- Advanced results handling
  - Concordances, collocations, counts, statistics …
- The corpus is available on the web

Linguistic contents and numbers

- The corpus contains dialect data from the national languages Danish, Faroese, Icelandic, Norwegian, and Swedish
- Contains speech data from app. 821 informants with 2.8 mill. words
- All the recordings represent spontaneous speech
- Differences in data collection due to differences in financing
  - Norwegian, Oevdalian, and some Danish: two kinds of recordings per informant:
    - semi-formal interview
    - informal conversation between two informants
  - Recordings of both young and old informants, both genders
  - Both new and old recordings
  - Audio or both audio and video recordings

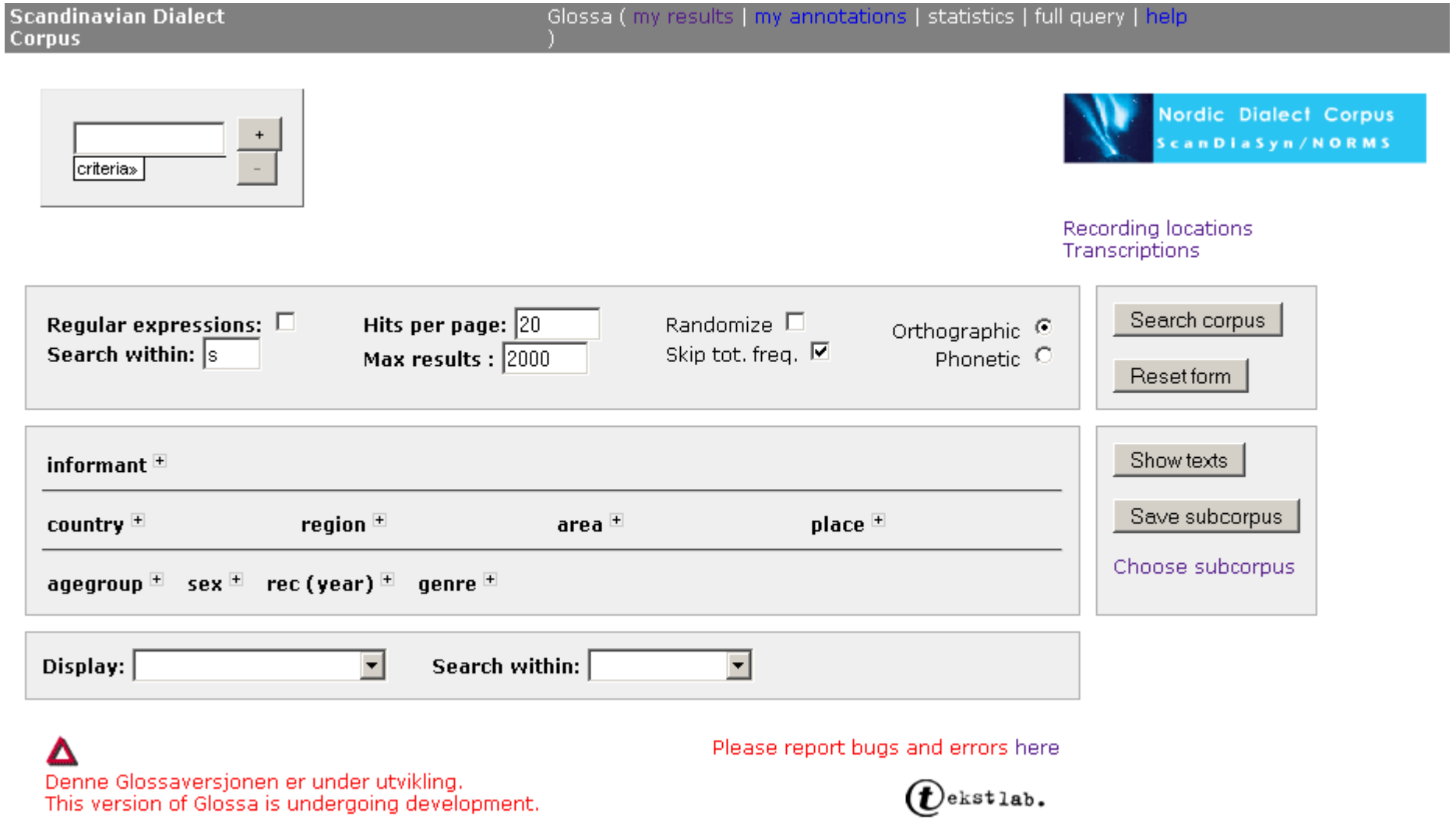
Annotation: transcription

- Each dialect has been transcribed by the standard official orthography of that country
  - In addition all the Norwegian dialects and some Swedish dialects have also been transcribed phonetically
    - The Norwegian phonetic transcription follows roughly that of Papazian and Helleland (2005). The transcription of the Oevdalian dialect follows the Oevdalian orthography (standardised in 2005 by the *Råðdjärum* – The Oevdalian Language Council).
  - The phonetic transcription is translated to an orthographic transcription via a semi-automatic dialect transliterator

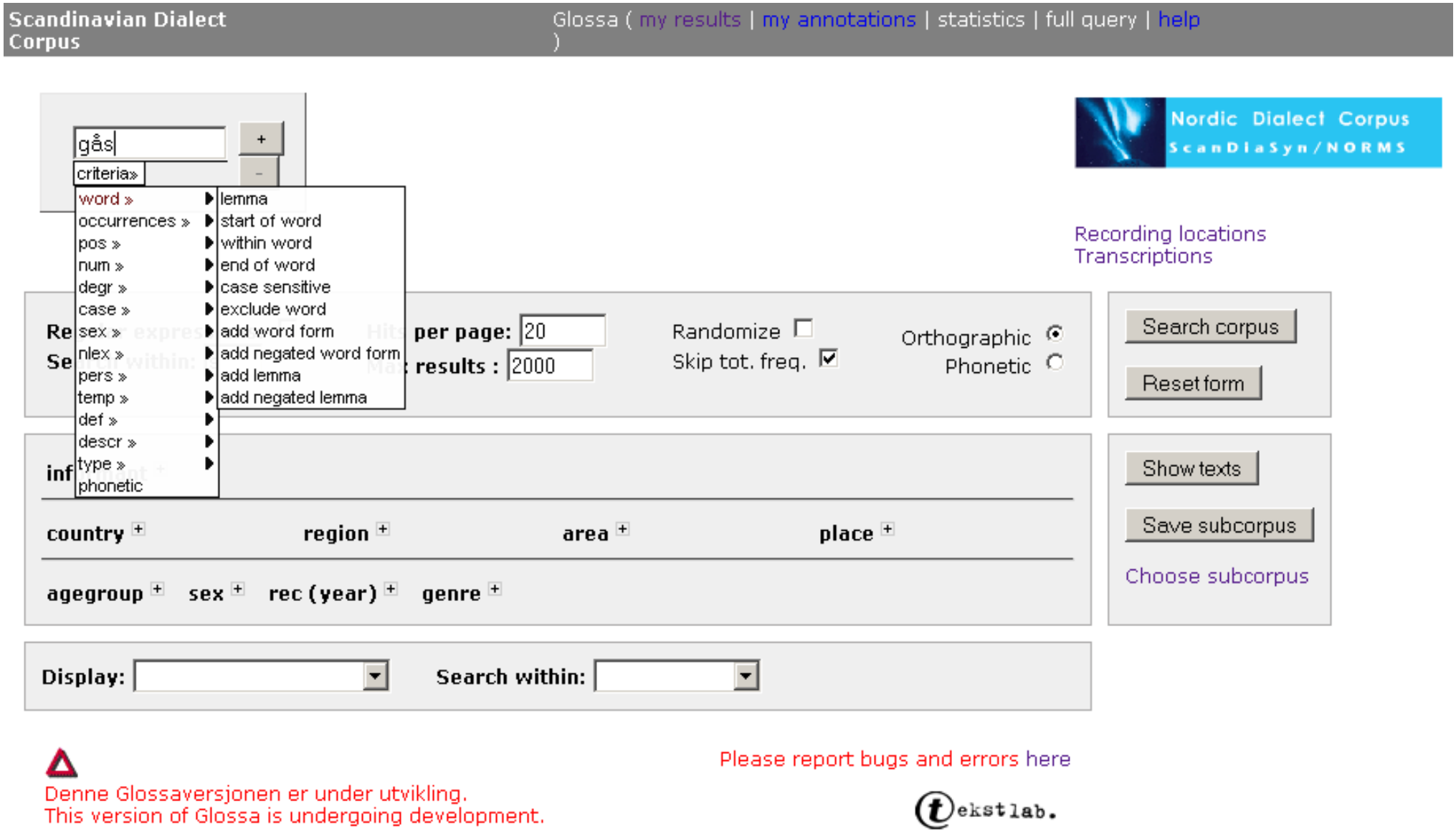
Annotation: tagging

- The corpus is POS tagged, with selected morpho-syntactic features language by language

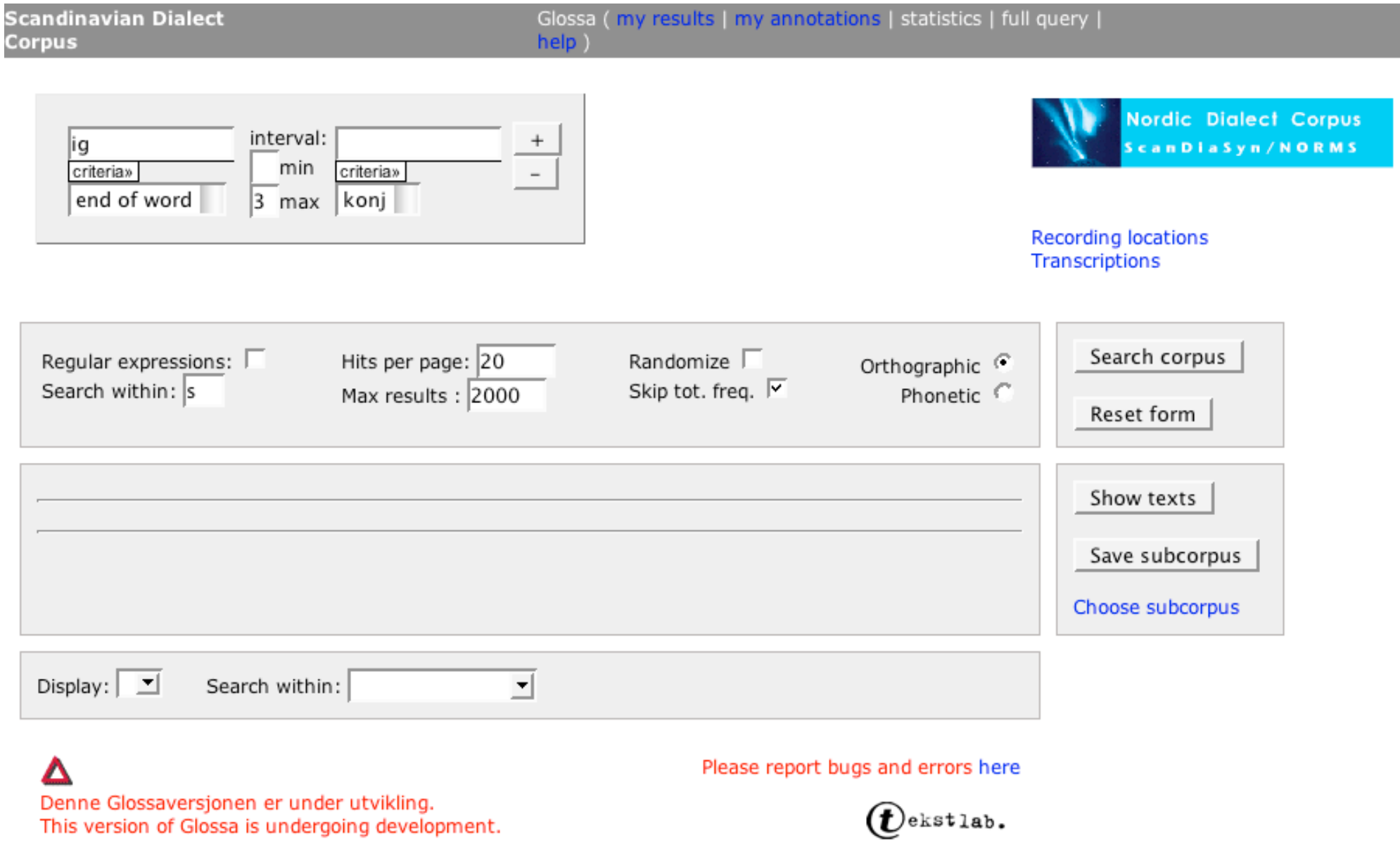
Search interface – Glossa



Searching for lemmas



Searching for more than one word



Search results

CWB expression: "([((word=".*ig" %c))] []{0,3} [((pos="konj"))]) ;"

Informants: 240
scandiasyn:
CWB expression: "([((word=".*ig" %c))] []{0,3} [((pos="konj"))]) ;"
Action :
: 474
Results pages: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

alvdal_02 påbygning # eller det er ikke så **veldig herlig det men**

alvdal_01 det er godt å bli **ferdig snart da ja men** (uforståelig) jævlig mye å gjøre i år

alvdal_01 ikke sett n # så nå har jeg jo **plutselig fire sjøl da for** jeg hadde jo tre # tre sjø

alvdal_02 tjener ikke så **jævlig mye men** # det blir jo litt

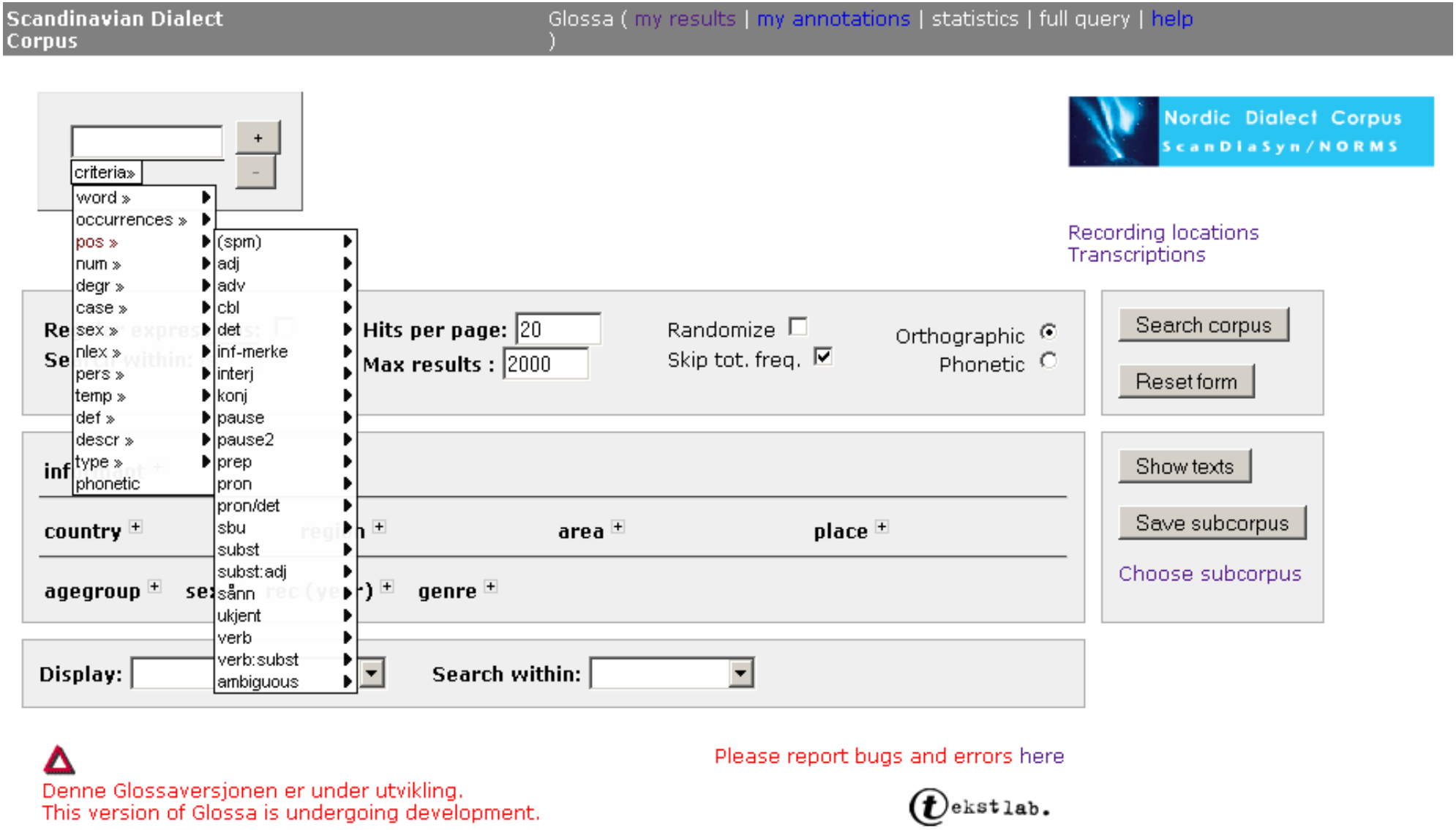
alvdal_01 ja det s- ja synes jeg ser så **herlig ut * men** jeg ...

alvdal_02 * m * yes * (latter) * det ser **herlig ut ja men** det er vel ett nålhue å komme gj

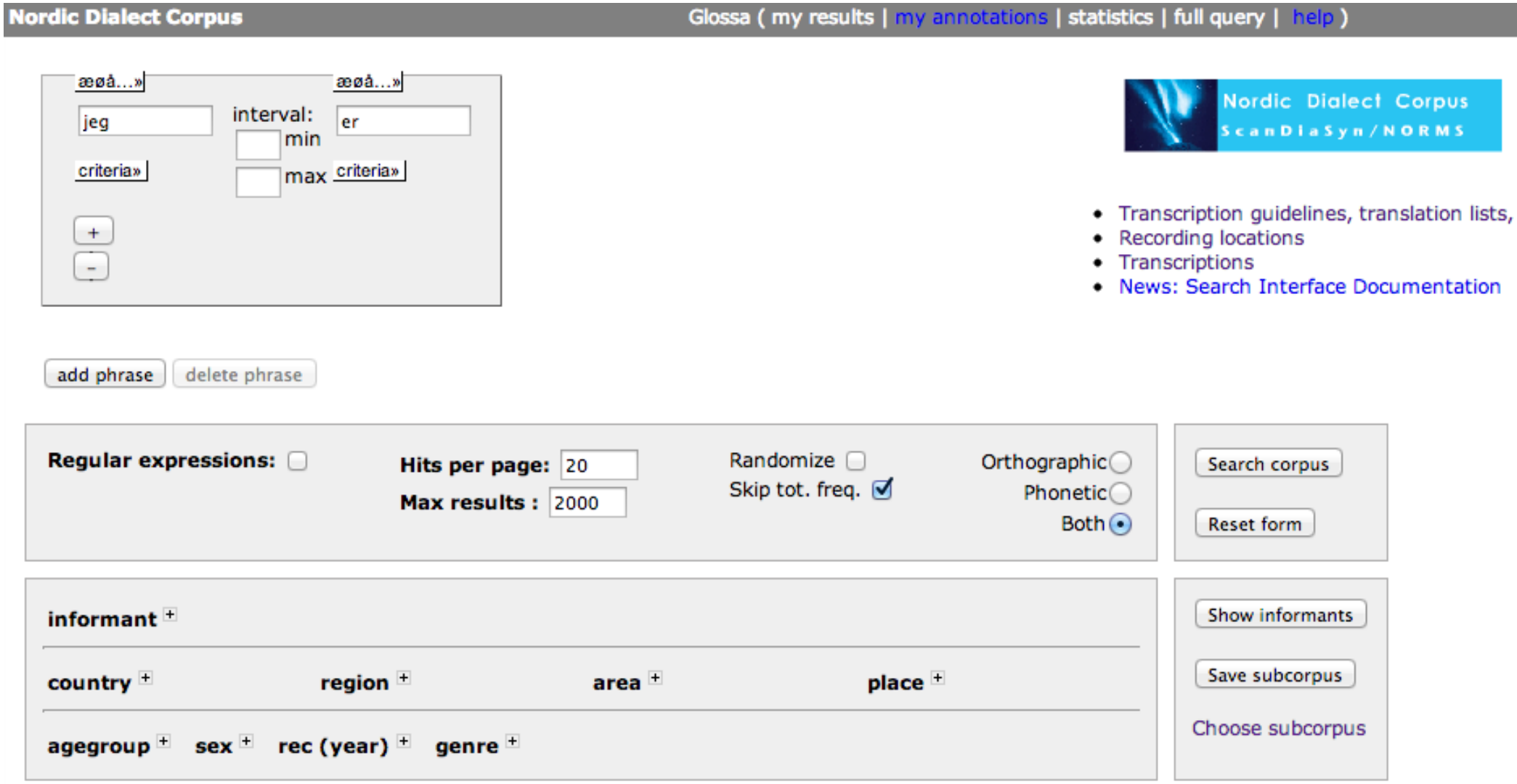
Some results presented as frequency list

occurences	match
8	**artig og**
6	**rolig og**
6	**egentlig men**
3	**mulig og**
3	**vanskelig og**
3	**egentlig # men**
3	**viktig og**
3	**mulig men**
2	**tidlig for**
2	**veldig kjekt og**
2	**frodig og**
2	**veldig fin oppvekst og**
2	**tidlig e begynne og**
2	**vanlig skole og**
2	**artig det og**
2	**veldig men**
2	**egentlig og**
2	**egentlig heile # oppveksten og**
2	**forferdelig råhølja heter det og**
2	**veldig gode venner i_hvert_fall og**
2	**veldig fin natur og**

Searching for part of speech



Phonetic querying



Displaying results

aasen_35 ja var det inte det ?

aasen_48 **ojojoj och hur fin (uforståelig) tänk att i går # så pratade M1 och jag just om ... # för då hade vi bröllopsdag i går**

aasen_35 jaha ?

Trouble viewing video?

context

± -

- Start +

- Stop +

>>

Informants: 240
scandiasyn:
CWB expression: "([(phon="um.*" %c)])] ;"
Action :
: 242
Results pages: 1 2 3 4 5 6 7 8 9 10 11 12 13

aasen_48	ojojoj och hur fin (uforståelig) tänk att i går # så pratade M1 och jag just **om** ... # för då hade vi bröllopsdag i går
ls	men vad gjorde **om** om rasterna ?
ls	men vad gjorde om **om** rasterna ?
aasen_35	och eh # till slut var det någon som var in där och tit- skulle se efter **om** hon hade haft varit darinne och där stod en slik # stor # brödkartong
ls	var det göra det_där byta **om** ? lära sig något nytt ?
aasen_35	det det var ju så det var ju många # stockholmare som var här **om** sommaren # över i gårdarna och hade ju lite # lärde ju oss lite av dem
aasen_35	ja det var verkligen riktig bocksvenska **om** de förstod det vet jag inte

- Phonetic and orthographic transcription

aal_01um nå skal vi nå er **jeg er** med i em (front-click) det derre prosjektet som H

no ska me no e **e e** me i em _ de dære prosjekkte somm HallingdaL Gå

Now we are going now **I'm** in em (front-click) the derre project Hallingdal G

aal_01um **jeg er** så glad i Al jeg

e e så gLa i ÅL e

I'm so happy I Eel (google)

aal_01um **jeg er** så en altså d- # jeg budde jo i Oslo ett år i_fjor

e e så n asså d # e budd jo i Ossjlo æitt år i_fjoL

I saw an ie d ... I lived in the Oslo one year i_fjor (google)

Display of transcription and tagging

CWB expression: **"([((word="bil.*" %c))]) ;"**
Action :
Hits found: **124**
Results pages: 1 **2** 3 4 5 6 7

alvdal_02 så skal jeg ta førerkort på bil nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre **bil**

alvdal_02 ja jeg # kjøpte meg **bil** i sommer # som jeg har pussa opp att

alvdal_02 har du traktor så # eller **bil** så kjører du heller det enn å sparke # går litt fortere

alvdal_04 nei da så er da mye men nå er det vel så du kommer ikke innover med **bil** heller nå vil jeg tru

alvdal_03 nye veien utover er så stille og så utoverbakke det er så at **bilene** bare trille utover jeg hører dem ikke i det hele tatt

alvdal_04 ja # ja når du skal ut med **bil** så er det vel det

lemma: bil
phon: bilan
pos: subst
sex: mask
num: fl
type: appell
defn: be
descr: me
nlex: 100

alvdal_03 men sa at jeg tar da ikke **bil** for å reise på skitur # hvis jeg i … 51

alvdal_03 og nå har jeg blitt for lat så jeg gidder ikke å gå opp her heller … kjøre på Klebersvanga det er helst der jeg kjører for jeg tør ikke kjøre opp

alvdal_03 og jeg tenkte at jeg kommer aldri hjem att med denne derre de… er

alvdal_03 og e så var jeg en ta disse postfolkene så da # " har du fått pr… **bilen** ? "

alvdal_03 " ja " sa " jeg har gjort det " men e ja vi har skulle ringe på n M… en det er ikke han som er # e der jeg går med **bilen** " sa jeg # men det v…

alvdal_03 ja jeg hadde ikke lovt han noe men han ville gjerne ha **bilen**

alvdal_04 ja dem er alldeles sjuk på gamle **biler** den # er m hun ene der a F9 hun kjøpte denne gamle Ladaen som jeg hadde vet du jeg hadde hadde Lada jeg

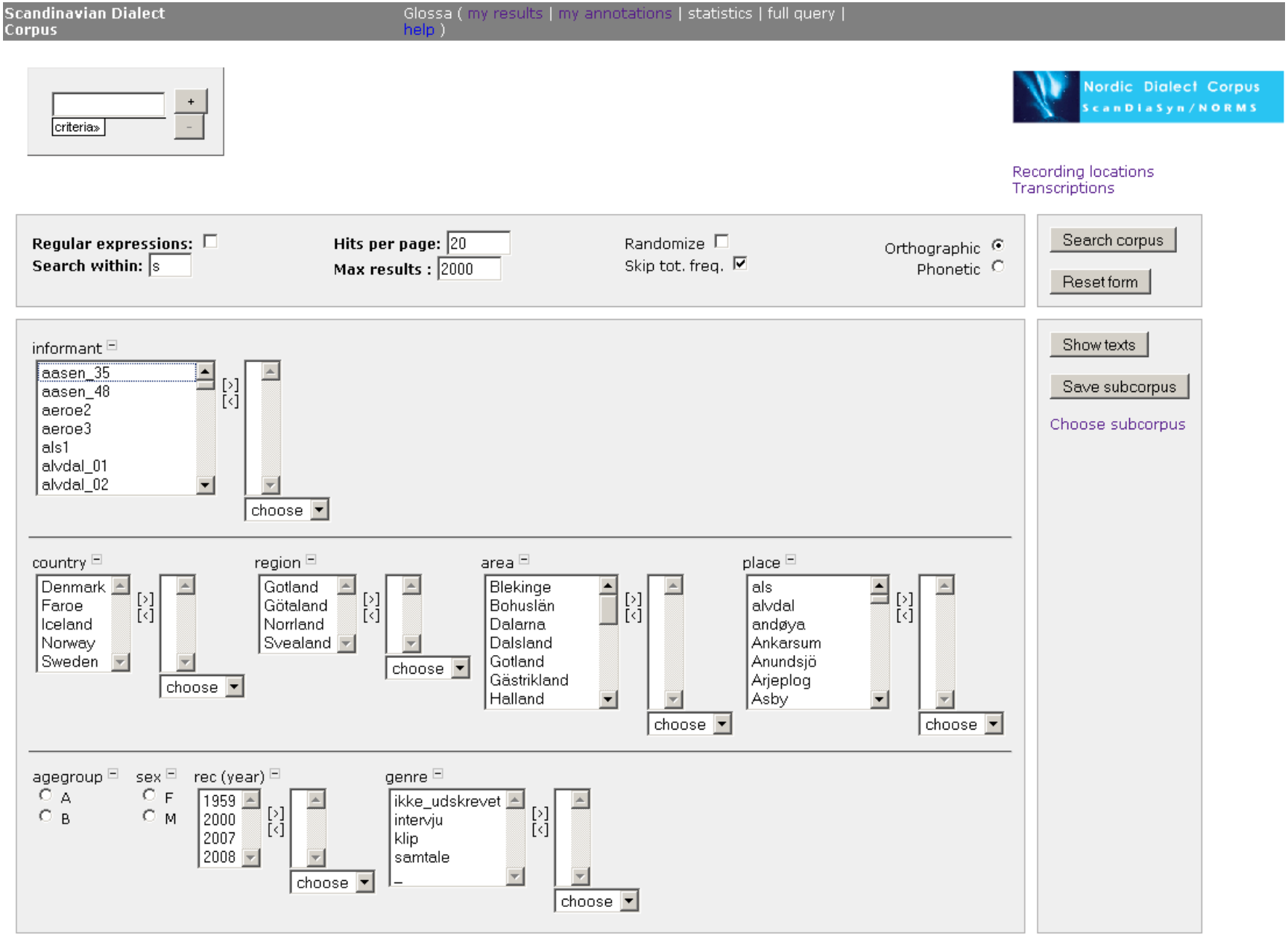
alvdal_04 men så byn- ja ja en gikk den men hun var alldeles sjuk på n vet du denne **bilen** så hun # kjøpte nå for en billig penge der

alvdal_04 men så byn- ja ja en gikk den men hun var alldeles sjuk på n vet du denne bilen så hun # kjøpte nå for en **billig** penge der

alvdal_03 hva_for en **bil** du har nå ?

alvdal_04 han driver på med **bilforretning** inni Oslo # så han skaffa meg denne der er en totusenmodell så den er det er en svær # firehjulstrekker og

Informant-based querying



Display information on informants 1

CWB expression: **"([((word="bil.*" %c))]) ;"**
Action :
Hits found: **124**
Results pages: 1 **2** 3 4 5 6 7

alvdal_02 så skal je
alvdal_02 ja jeg # k
alvdal_02 har du tra
alvdal_04 nei da så
alvdal_03 nye veien
alvdal_04 ja # ja nå
alvdal_03 men sa at
alvdal_03 og nå har
alvdal_03 og jeg ter
alvdal_03 og e så v
alvdal_03 " ja " sa "
alvdal_03 ja jeg hadde ikke lovt han noe men han ville gjerne ha **bilen**
alvdal_04 ja dem er alldeles sjuk på gamle **biler** den # er m hun ene der a F9 hun kjøpte denne gamle Ladaen som jeg hadde vet du jeg
alvdal_04 men så byn- ja ja en gikk den men hun var alldeles sjuk på n vet du denne **bilen** så hun # kjøpte nå for en billig penge der
alvdal_04 men så byn- ja ja en gikk den men hun var alldeles sjuk på n vet du denne bilen så hun # kjøpte nå for en **billig** penge der
alvdal_03 hva_for en **bil** du har nå ?

nele tatt
t der jeg kjører fc
jeg går med **bile**

Mozilla Firefox

http://omilia.uio.no/glossa//html/profile.php?tid=alvdal_03&corpus=scandiasyn

Informant details for *alvdal_03* in the Scandiasyn corpus

SELECT tid,sex,agegroup,country,region,area,place,wc,rec FROM SCANDIASYNauthor WHERE tid = 'alvdal_03'

Code	Sex	Age group	Country	Region	Area	Place	Word count	Recorded
alvdal_03	F	B	Norway			alvdal	4876	2008

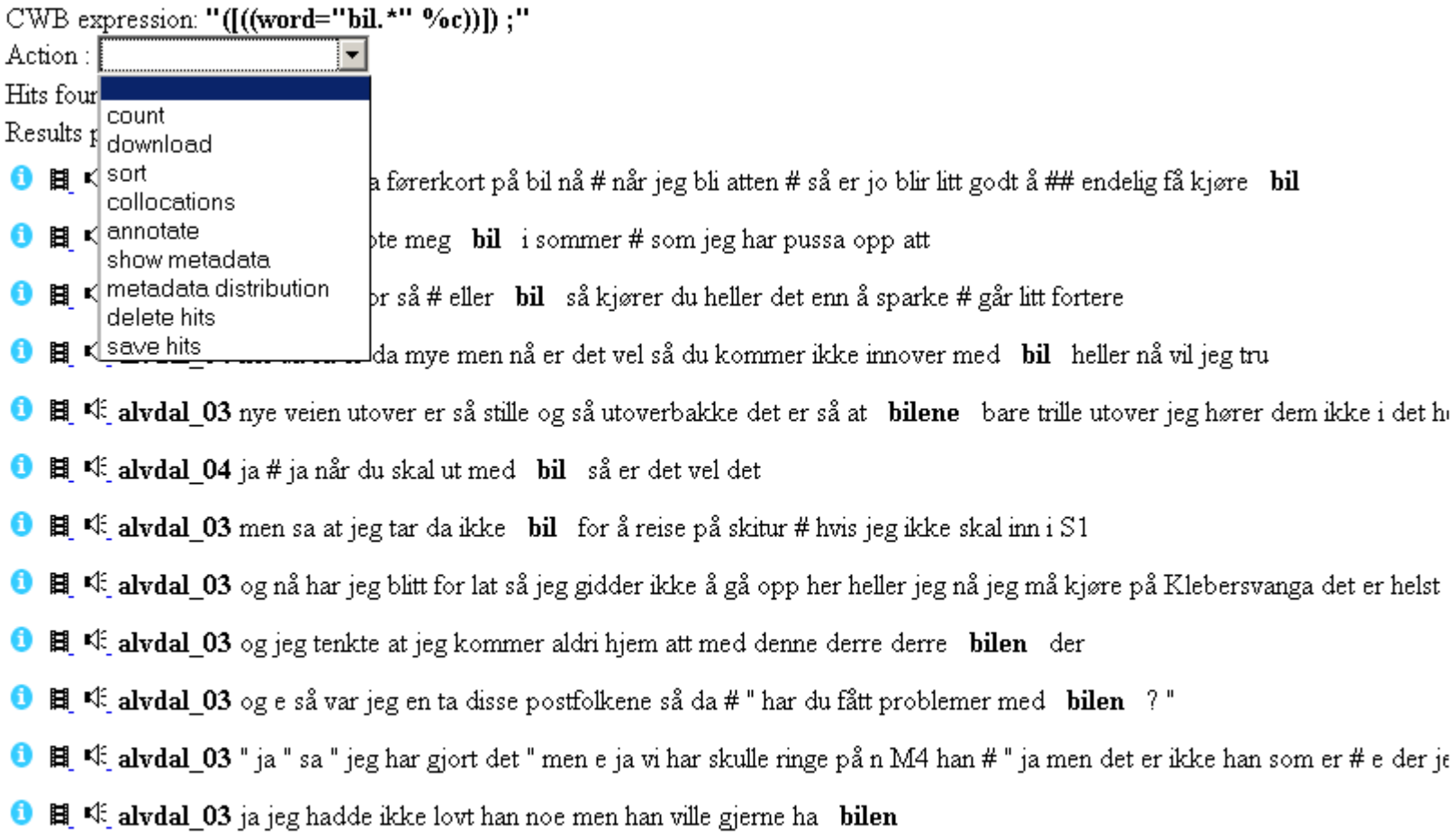
Fullført

Display information on informants 2

SELECT SUM(wc) FROM SCANDIASYNauthor WHERE active REGEXP '^(y)$' AND active IS NO
Word count for selected informants: 2824585 (total for Scandiasyn corpus: 2824585)
Selected informants: 821 from 228 places in 5 countries

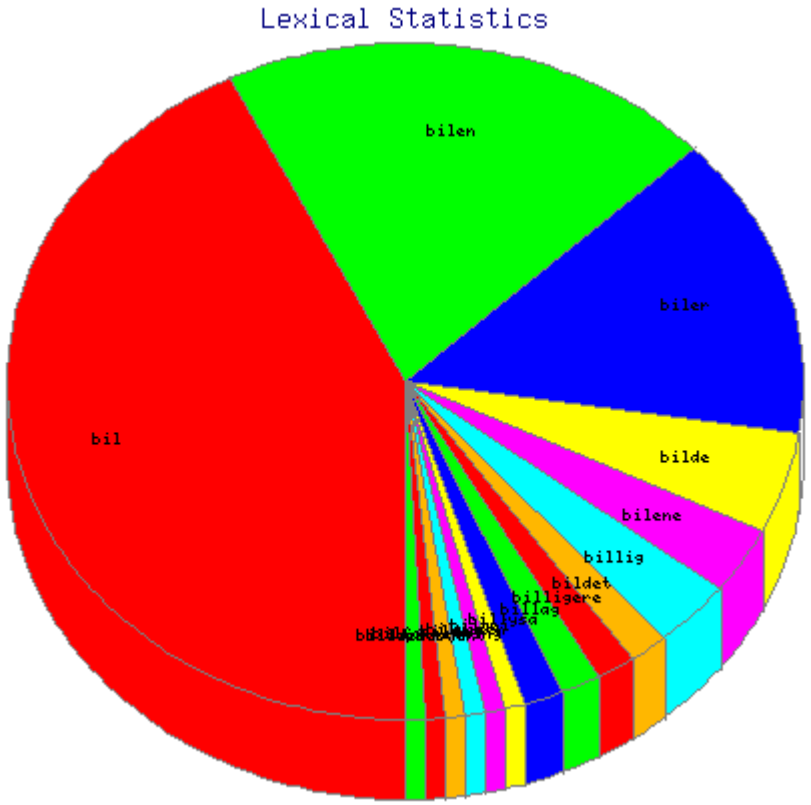
Code	Sex	Age group	Country
ankarsrum_om1	M	B	Sweden
ankarsrum_om3	M	B	Sweden
ankarsrum_ym1	M	A	Sweden
anundsjo_om2	M	B	Sweden
anundsjo_ow3	F	B	Sweden
arjeplog_om1	M	B	Sweden
arsunda_ow1	F	B	Sweden
arsunda_om1	M	B	Sweden
arsunda_om2	M	B	Sweden
arsunda_ow2	F	B	Sweden
asby_om2	M	B	Sweden
asby_om3	M	B	Sweden

Action menu


CWB expression: "([((word="bil.*" %c))]) ;"
Action :
Hits four
Results p
count
download
sort
collocations
annotate
show metadata
metadata distribution
delete hits
save hits
a førerkort på bil nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre bil
te meg bil i sommer # som jeg har pussa opp att
r så # eller bil så kjører du heller det enn å sparke # går litt fortere
da mye men nå er det vel så du kommer ikke innover med bil heller nå vil jeg tru
alvdal_03 nye veien utover er så stille og så utoverbakke det er så at bilene bare trille utover jeg hører dem ikke i det h
alvdal_04 ja # ja når du skal ut med bil så er det vel det
alvdal_03 men sa at jeg tar da ikke bil for å reise på skitur # hvis jeg ikke skal inn i S1
alvdal_03 og nå har jeg blitt for lat så jeg gidder ikke å gå opp her heller jeg nå jeg må kjøre på Klebersvanga det er helst
alvdal_03 og jeg tenkte at jeg kommer aldri hjem att med denne derre derre bilen der
alvdal_03 og e så var jeg en ta disse postfolkene så da # " har du fått problemer med bilen ? "
alvdal_03 " ja " sa " jeg har gjort det " men e ja vi har skulle ringe på n M4 han # " ja men det er ikke han som er # e der je
alvdal_03 ja jeg hadde ikke lovt han noe men han ville gjerne ha bilen

Count

occurences	match
53	bil
25	bilen
18	biler
6	bilde
4	bilene
4	billig
2	bildet
2	billigere
2	billag
2	billysa
1	bilvei
1	bilveien
1	bilveg
1	bilforretning
1	bildeprosjekt
1	bildeler



Deleting or selecting some results

CWB expression: **"([(word="bil.*" %c)]) ;"**
Action :
Hits found: **124**
Results pages: 1 **2** 3 4 5 6 7

Delete selection | select all | unselect all | Finished deleting

- **alvdal_02** så skal jeg ta førerkort på bil nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre **bil**
- **alvdal_02** ja jeg # kjøpte meg **bil** i sommer # som jeg har pussa opp att
- **alvdal_02** har du traktor så # eller **bil** så kjører du heller det enn å sparke # går litt fortere
- **alvdal_04** nei da så er da mye men nå er det vel så du kommer ikke innover med **bil** heller nå vil jeg tru
- **alvdal_03** nye veien utover er så stille og så utoverbakke det er så at **bilene** bare trille utover jeg hører dem ikke i det hele tatt
- **alvdal_04** ja # ja når du skal ut med **bil** så er det vel det
- **alvdal_03** men sa at jeg tar da ikke **bil** for å reise på skitur # hvis jeg ikke skal inn i S1
- **alvdal_03** og nå har jeg blitt for lat så jeg gidder ikke å gå opp her heller jeg nå jeg må kjøre på Klebersvanga det er helst der jeg (uforståelig)
- **alvdal_03** og jeg tenkte at jeg kommer aldri hjem att med denne derre derre **bilen** der
- **alvdal_03** og e så var jeg en ta disse postfolkene så da # " har du fått problemer med **bilen** ? "

Annotating results

CWB expression: **"([((word="hun" %c))][((pos="subst"))]) ;"**
Action :
Hits found: **96**
Results pages: 1 2 3 4 5

Save annotations (* indicates: no reviewed value stored.)

pronoun *

mefjordvaer_19 ja men det hender jo at # at vi slår til og så bruker uttrykk og d- **hun datter** vår som bor i # Oslo hun em # hun jol

pronoun *

mefjordvaer_19 " ja hvor fikk **hun tak** i de jernene ? " " jo det var smeden i Været " sa han # " som laga jernet til hun " sa han

demonstrative *

stange_03 og e # ja jeg vet ikke åssen jeg skal si det # nei jeg har akkurat e s- sagt åt **hun dama** de skrev at det jeg kan tenk

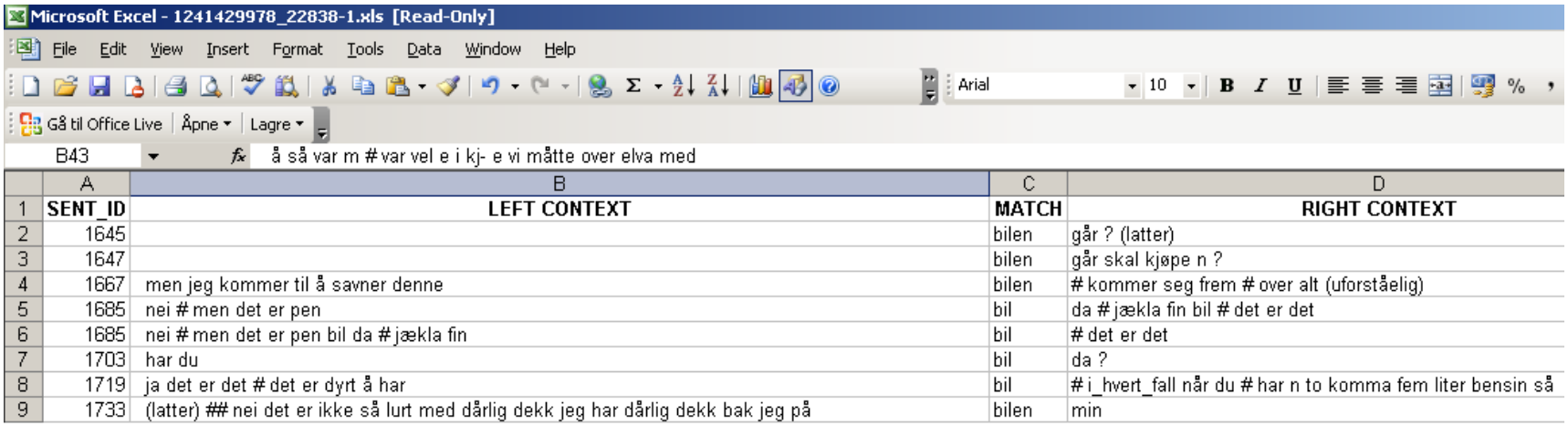
pronoun *

stonglandseide_47 når du kommer dit som du skulle tømme henne så har jo **hun hålt** lekke tom på veien og sånn blir det med skatten veien

pronoun *

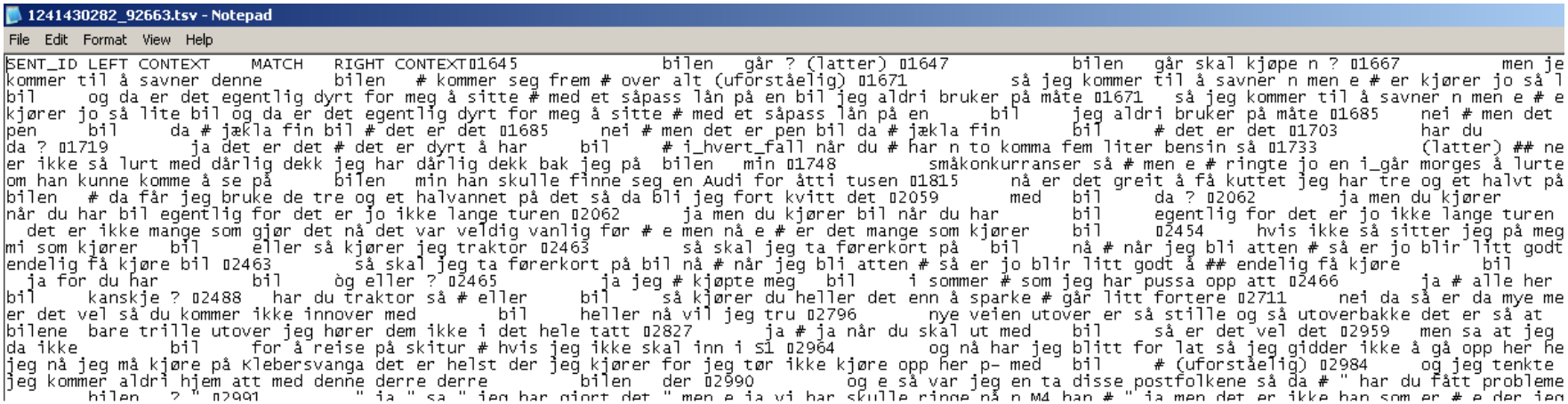
Downloading results, examples:

Excel:



	A	B	C	D
1	SENT_ID	LEFT CONTEXT	MATCH	RIGHT CONTEXT
2	1645		bilen	går ? (latter)
3	1647		bilen	går skal kjøpe n ?
4	1667	men jeg kommer til å savner denne	bilen	# kommer seg frem # over alt (uforståelig)
5	1685	nei # men det er pen	bil	da # jækla fin bil # det er det
6	1685	nei # men det er pen bil da # jækla fin	bil	# det er det
7	1703	har du	bil	da ?
8	1719	ja det er det # det er dyrt å har	bil	# i_hvert_fall når du # har n to komma fem liter bensin så
9	1733	(latter) ## nei det er ikke så lurt med dårlig dekk jeg har dårlig dekk bak jeg på	bilen	min

Tab separated values:




```
SENT_ID LEFT CONTEXT    MATCH   RIGHT CONTEXT01645               bilen   går ? (latter) 01647            bilen   går skal kjøpe n ? 01667          men je
kommer til å savner denne       bilen   # kommer seg frem # over alt (uforståelig) 01671          så jeg kommer til å savner n men e # er kjører jo så l
bil     og da er det egentlig dyrt for meg å sitte # med et såpass lån på en bil jeg aldri bruker på måte 01671    så jeg kommer til å savner n men e # e
kjører jo så lite bil og da er det egentlig dyrt for meg å sitte # med et såpass lån på en      bil     jeg aldri bruker på måte 01685    nei # men det
pen     bil     da # jækla fin bil # det er det 01685     nei # men det er pen bil da # jækla fin      bil     # det er det 01703          har du
da ? 01719        ja det er det # det er dyrt å har     bil     # i_hvert_fall når du # har n to komma fem liter bensin så 01733          (latter) ## ne
er ikke så lurt med dårlig dekk jeg har dårlig dekk bak jeg på  bilen   min 01748          småkonkurranser så # men e # ringte jo en i_går morges å lurte
om han kunne komme å se på      bilen   min han skulle finne seg en Audi for åtti tusen 01815     nå er det greit å få kuttet jeg har tre og et halvt på
bilen   # da får jeg bruke de tre og et halvannet på det så da bli jeg fort kvitt det 02059      med   bil     da ? 02062        ja men du kjører
når du har bil egentlig for det er jo ikke lange turen 02062      ja men du kjører bil når du har       bil     egentlig for det er jo ikke lange turen
  det er ikke mange som gjør det nå det var veldig vanlig før # e men nå e # er det mange som kjører    bil     02454     hvis ikke så sitter jeg på meg
mi som kjører   bil     eller så kjører jeg traktor 02463         så skal jeg ta førerkort på   bil     nå # når jeg bli atten # så er jo blir litt godt
endelig få kjøre bil 02463        så skal jeg ta førerkort på bil nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre        bil
  ja for du har         bil     òg eller ? 02465          ja jeg # kjøpte meg   bil     i sommer # som jeg har pussa opp att 02466        ja # alle her
bil     kanskje ? 02488   har du traktor så # eller     bil     så kjører du heller det enn å sparke # går litt fortere 02711     nei da så er da mye me
er det vel så du kommer ikke innover med        bil     heller nå vil jeg tru 02796       nye veien utover er så stille og så utoverbakke det er så at
bilene  bare trille utover jeg hører dem ikke i det hele tatt 02827       ja # ja når du skal ut med    bil     så er det vel det 02959   men sa at jeg
da ikke         bil     for å reise på skitur # hvis jeg ikke skal inn i si 02964         og nå har jeg blitt for lat så jeg gidder ikke å gå opp her he
jeg nå jeg må kjøre på Klebersvanga det er helst der jeg kjører for jeg tør ikke kjøre opp her p- med   bil     # (uforståelig) 02984     og jeg tenkte
jeg kommer aldri hjem att med denne derre derre         bilen   der 02990         og e så var jeg en ta disse postfolkene så da # " har du fått probleme
        bilen   ? " 02991         " ja " sa " jeg har gjort det " men e ja vi har skulle ringe nå n M4 han # " ja men det er ikke han som er # e der jeg
```

Saving results

CWB expression: **"([(lemma="bil" %c)]) ;"**

Action :

Hits four

Results p

- count
- download
- sort
- collocations
- annotate
- show metadata
- metadata distribution
- delete hits
- save hits

? (latter)

mer til å savner denne **bilen** # kommer seg frem # over alt (uforståelig)

er til å savner n men e # er kjører jo så lite **bil** og da er det egentlig dyrt for meg å sitte # med et såpass lån på en bil jeg aldri bruker på måte

er til å savner n men e # er kjører jo så lite bil og da er det egentlig dyrt for meg å sitte # med et såpass lån på en **bil** jeg aldri bruker på måte

alvdal_02 ja det er det # det er dyrt å har **bil** # i_hvert_fall når du # har n to komma fem liter bensin så

alvdal_01 småkonkurranser så # men e # ringte jo en i_går morges å lurte på om han kunne komme å se på **bilen** min han skulle finne seg en Audi for åtti tusen

alvdal_01 ja men du kjører **bil** når du har bil egentlig for det er jo ikke lange turen

alvdal_01 ja men du kjører bil når du har **bil** egentlig for det er jo ikke lange turen

alvdal_01 det er ikke mange som gjør det nå det var veldig vanlig før # e men nå e # er det mange som kjører **bil**

alvdal_02 hvis ikke så sitter jeg på meg mor mi som kjører **bil** eller så kjører jeg traktor

alvdal_02 så skal jeg ta førerkort på **bil** nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre bil

alvdal_02 så skal jeg ta førerkort på bil nå # når jeg bli atten # så er jo blir litt godt å ## endelig få kjøre **bil**

How to get several transcriptions quickly

Use a semi-automatic transliteration program

Input: a phonetically transcribed file
Processing:

- Pick a similar dialect as a starting point
- Transliterate to standard orthography, then train. Continue to transliterate, train again.

Ouput: aligned phonetic and orthographic transcriptions

Dialekttranslitterator

Transkripsjoner

Ny transkripsjon | Ny dialekt | Slett valgte

Dialekt	Transkripsjonsfil	Dato	Last ned
roeros	roeros_04gk-amr_ifg_kk.trs	04.02.2011	
kirkenes	kirkenes_01um-kb_hl_kk.trs	17.02.2011	
kirkenes	kirkenes_02uk-kb_hl_kk.trs	17.02.2011	
bjugn	bjugn_15-19_sb_ifg.trs	17.02.2011	
bjugn	bjugn_15-aml_sb_ifg.trs	22.02.2011	
bjugn	bjugn_16-ma_sr_kk.trs	22.02.2011	
bjugn	bjugn_19-aml_sb_ifg.trs	22.02.2011	
kirkenes	kirkenes_03gm-kb_kk_lh.trs	24.02.2011	
kirkenes	kirkenes_04gk-kb_kk_lh.trs	24.02.2011	
bjugn	bjugn_23-pmk_hl_kk.trs	24.02.2011	
roeros	roeros_03gm-04gk_ifg.trs	24.02.2011	
kjoellefjord	kjoellefjord_03gm-04gk_sb_amg_lh.trs	28.02.2011	
beiarn	beiarn_01um-kb_kk_sb.trs	01.03.2011	
beiarn	beiarn_02uk-kb_kk_sb.trs	03.03.2011	
beiarn	beiarn_01um-02uk_kk_sb.trs	03.03.2011	
Oppdal	oppdal_02-31_eo_hl_ifg.trs	11.03.2011	
beiarn	beiarn_03gm-kb_sb_kk.trs	17.03.2011	
beiarn	beiarn_04gk-kb_sb_kk.trs	17.03.2011	
beiarn	beiarn_03gm-04gk_kk_sb.trs	17.03.2011	
lavangen	lavangen_03gm-ov_sb_eo.trs	17.03.2011	
lavangen	lavangen_04gk-ov_sb_eo.trs	19.03.2011	
lavangen	lavangen_03gm-04gk_sb_eo.trs	19.03.2011	
skaugdalen	skaugdalen_36-43_hl_kk_ifg.trs	21.03.2011	
vardoe	vardoe_01um-kb_eo_kk.trs	23.03.2011	
vardoe	vardoe_02uk-kb_eo_kk.trs	23.03.2011	
vardoe	vardoe_01um-02uk_eo_kk.trs	24.03.2011	
heroeyMR	heroeyMR_01um-ta_hl_lh.trs	28.03.2011	
heroeyMR	heroeyMR_02uk-ta_hl_lh.trs	29.03.2011	
heroeyMR	heroeyMR_01um-02uk_hl_lh.trs	29.03.2011	
Stranda	stranda_01um-ta_hl_kk.trs	30.03.2011	

Fildeler

Filnavn	Sjekket?
kirkenes_01um-kb_hl_kk.trs.01	
kirkenes_01um-kb_hl_kk.trs.02	
kirkenes_01um-kb_hl_kk.trs.03	
kirkenes_01um-kb_hl_kk.trs.04	
kirkenes_01um-kb_hl_kk.trs.05	
kirkenes_01um-kb_hl_kk.trs.06	

Sett backoff

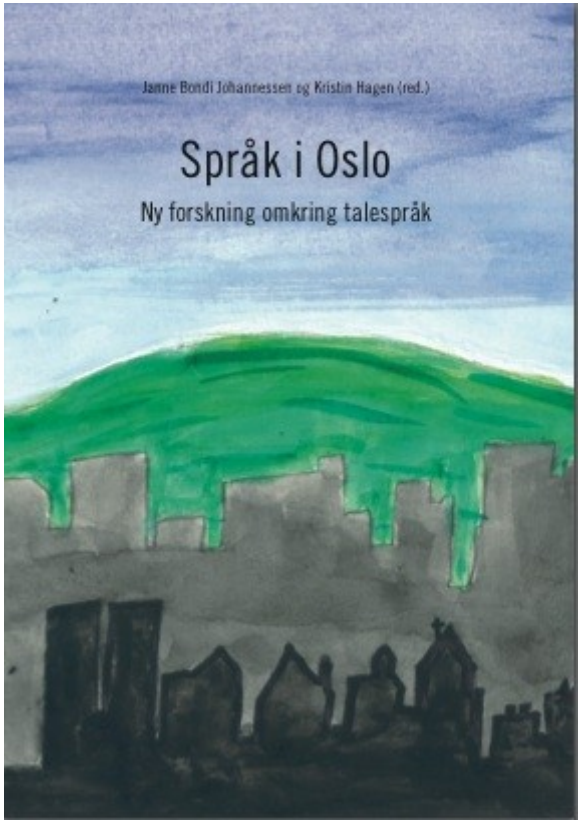
0.01	aal
0.01	aamot
0.01	aaseral
0.01	aasnes
0.01	alvdal
0.2	andoeya
0.01	aremark
0.01	aukra
0.01	aure
0.01	aurland
0.2	Ballangen
0.01	bardu
0.01	bergen
0.01	bjugn
0.5	bodoe
0.01	boe
0.01	boemlo
0.01	botnhavn
0.01	brandbu
0.01	brekkom
0.01	brunlanes
0.01	bud
0.01	bykle
0.01	dalsbygda

"ja"	"ja"	1
"je	"jeg	1
"je"	"jeg"	1
"jentene"	"jentene"	1
"jo	"jo	1
"jæi"	"jeg"	1
"ka	"hva	1
"ker	"hvor_	2
"kerr	"hvor_	4
"kerr"	"hvor_"	4
"kær	"hvor_	1
"kærr	"hvor_	2
"kærr"	"hvor_"	1
"kæssj"	"hvor_"	1

Result of transliteration:
a bilingual dictionary

- Nordic dialect corpus:
- http://omilia.uio.no/glossa/html/index_dev.php?corpus=
- scandiasyn

Book:



- Janne Bondi Johannessen and Kristin Hagen (eds.) 2008. *Language in Oslo – New research on spoken language.* Oslo: Novus.

Ruth Vatvedt Fjeld (UiO):

Talespråksforskningens betydning for leksikografien

- Talespråksord mangler i ordbøkene:
  - *Herre*, men ikke *derre*
  - Ikke *loka, lættis*
  - toalettpapir, klosettpapir, men ikke dasspapir, dassrull

- Betydningsspesialisering
- --- det gjør jo automatisk at hvis alt går i dass da så --
- --- oi jeg sølte (uforståelig) # (stønning) # jeg er helt dass i øya

- Betydningsutvidelse
  - *skikkelig a2 (fra lty) bra, god, ordentlig ha s- e klær / få s- betaling / s-e folk / adv: oppføre seg s-*
  - den smakte skikkelig # forferdelig

Toril Opsahl, Unn Røyneland og Bente Ailin Svendsen (UiO):
"Syns du jallanorsk er lættis, eller?" – om taggen [lang=X] i NoTa-Oslo-korpuset

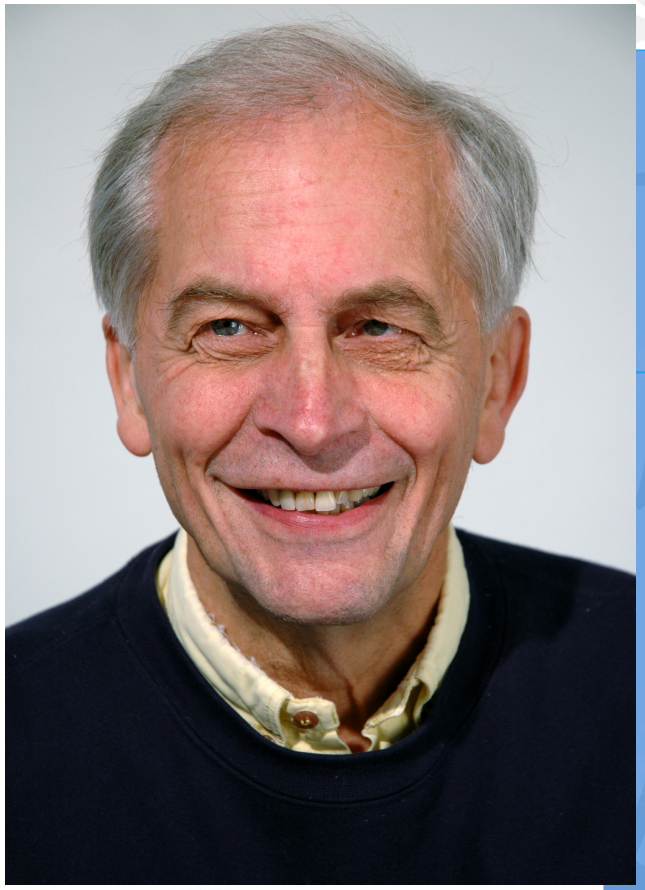


- Hva slags ord finnes i korpuset som ikke finnes i ordbøkene?
- Hvilke befolkningsgrupper bruker mest unormerte ord?
- Er det forskjeller?
- Gutter fra østre deler av Oslo er mest kreative i dette nyordsperspektivet.

Kjell Ivar Vannebo (UiO):
NoTa-informantene og tellemåten

- Gammel tellemåte: *treogtredve*
- Ny tellemåte: *trettitre*

- Det er over femti år siden den nye tellemåten ble innført.
  - 1950, enstemmig i Stortinget, intitiativ fra Telegrafverket
- Eneste gang en normering for talemål har funnet sted for norsk
- Hva kan man konkludere med grunnlag i NoTa-korpuset?
- Reformen har vært vellykket, selv om andre har hevdet det motsatte!
  - 18-25 år: 99 % ny tellemåte
  - 26-50 år: 49 %
  - 51+ år: 32 %



Øystein Alexander Vangsnes (UiT):

Omkring adnominalt åssen/hvordan *i Oslo-målet*

- *hvordan* poker var det du spilte da ?
- *åssen* bikkje er det men- ... ?

- *Åssen* er mer brukt i Oslo Rest enn i Oslo Vest
- Adnominalt *åssen* er vanligere enn adnominalt *hvordan*
- Adnominalt *åssen* er uvanlig blant folk med høy utdanning
- Betydningen til adnominalt åssen:
- Egenskap
  - Åssen bil kjøpte'ru'a? En rød en.
- Eksemplar
  - åssen topp ? ja e på ballen ... på toppen av ballen

Janne Bondi Johannessen (UiO):
Psykologiske demonstrativer

- jo jeg var jo på, på Ammerudhjemmet, og da så satt *hun dama* som jeg snakka om som var snart hundre år (Dame, 80, Grorud, Oslo)
- men hva med han derre m leste du om han derre *han tyskeren* som hadde kuttet av utstyret på en fyr og spist det (Mann, 18, Vestre Aker)

- En ny type demonstrativ
  - De vanlige (den, denne)
    - Geografisk avstand – nær, fjern
  - De nye (han, hun)
    - psykologisk distanse

Svein Lie (UiO):
Veldig sånn festejente

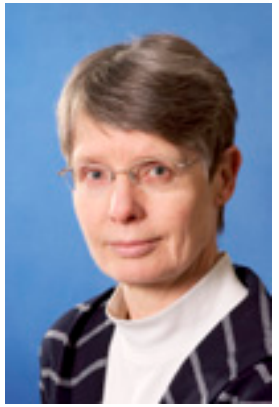
- Nye betydninger *av veldig* og *sånn*
- *veldig* brukt med substantiv
  - man oppfatter et bestemt trekk eller egenskap ved substantivet, som så kan graderes.
- *Sånn* – et ekstremt vanlig talespråksord, har mange viktige bruksbetingelser
  - å signalisere høflighet!

Gjert Kristoffersen (UiB) og Hanne Gram Simonsen (UiO):

Osjlo! En undersøkelse av sl-*sekvensen i NoTa-korpuset*

- Apikal, bakre, uttale eller fremre uttale av <sl>?
- 3477 belegg av <sl>
  - Fremlyd: over 99,6 % bakre uttale
  - Innlyd: 68 % bakre uttale
- Ordet *Oslo*:
  - 84,4 % bakre uttale (inkludert alle de unge)

UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET



Elisabet Engdahl (GU):
Frågor i NoTa

- Spørsmålsformuleringer
  - ja/nei-spørsmål
  - Ekkospørsmål
    - han sa måtte pappa måtte fikse det # for det han har installert det
    - installert hva da ? (Ekkospørsmål)
  - Spørsmål med deklarativ ordstilling
    - du har ikke vært der du heller nei ?
- bruksbetingelser
- NoTa-korpuset som forskningsverktøy
  - hvordan transkribørene har annotert spørsmålene
  - hvordan korpuset fungerer.

Inger Margrethe Hvenekilde Seim (UiO):

Innhold og struktur i en samtale mellom to ungdommer i et flerkulturelt miljø i Oslo

- To gutter med innvandrerbakgrunn i en timelange samtale.
  - Hvordan forløper samtalen?
  - Hvordan tas emner opp?
- Emnet etnisitet
  - Egen gruppetilhørighet
  - Hverandres gruppetilhørighet
  - Andres gruppetilhørighet
  - Akkurat dette emnet kommer guttene inn på flere ganger gjennom hele samtalen.

Jan Svennevig (BI/UiO):

"Ikke sant" som respons i samtale

- Sammenligning av flere små, eldre samtalekorpus og NoTa-korpuset
- *ikke sant* brukt som respons
  - har økt fra null i 1994 til en tredel av alle forekomster i 2005.
  - mest vanlig blant yngre mennesker.
- flere betydninger for uttrykket brukt som respons
  - felles for dem er at det brukes som bekreftelse på noe som er sagt i samtalen.
- Kan virke ganske irriterende på samtalepartneren, og Svennevig forklarer hvorfor.

Lars-Olof Delsing (LU):

Viskningar och rop – eller hur vi undrar och förundras

- Ordstillingen og ordene som er brukt i utrops- og undringskonstruksjoner.
  - Så du har vackra rosor! (f.svensk)
  - Så vakre roser du har! (norsk)
  - Vad/så många slipsar du har! (svensk)
  - Så mange slips du har! (norsk)
- En enorm variasjon mellom de skandinaviske språkene. Men likevel en del generaliseringer.

Gunnar Hrafn Hrafnbjargarson (UiO/UiT):

Substantiverte adjektiv: Det er verste jeg har hørt

- Nakne superlativer : En forholdsvis ukjent konstruksjonstype
- Svakt bøyd, men mangler artikkel,
- Ligner på utbryting, men er det ikke
  - det er bare fugler som kan fly
  - det er eneste som er minuset syns jeg da
- Dette er ikke utbrytinger, slik de tilsynelatende kan se ut
  - Det som kan fly er bare fugler
  - * Det som er minuset synes jeg er eneste
- men er faktisk en type ikke-referensiell, substantivert predikativ.

Marit Julien (LU):
Så vanleg at det kan ikkje avfeiast – om V2 i innføyde setningar.

- Ordstillingen i underordnede setninger
  - Men det som er er at *han kan ikke* lage sanger.
  - Så jeg bare sier at det kan jeg ikke gjøra.

- Hva slags betingelser er mulige?
- Det ikke spiller noen direkte rolle hva slags verb som innleder at-setningen.
- Det som teller, er utelukkende semantikk: at-setningen må være hevdet.
  - *Men jeg tviler på at slike konserter hjelper faktisk mot volden.

Mari Nygård, Kristin M. Eide og **Tor A. Åfarli** (NTNU):
Ellipsens syntaktiske struktur

UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

- Lydløse ord. Uuttalte, lydløse eller stumme ord. Ellipse.
  - __ fikk ikke lov å herje som vi ville.
  - nei jeg var vel en ti-tolv år tenker jeg, så __ var ikke så gammal ...
- Ikke noen allmenn forståelse blant språkforskerne om fenomenet ellipse.
  - Noen: det som ikke er uttalt, finnes heller ikke.
  - Nygård, Eide og Åfarli: ellipse er like eksisterende i språket som de ordene vi faktisk hører.
  - de tilsynelatende stumme ordene kan bare forekomme i klart definerte kontekster, og ikke er tilfeldig spredt rundt omkring.

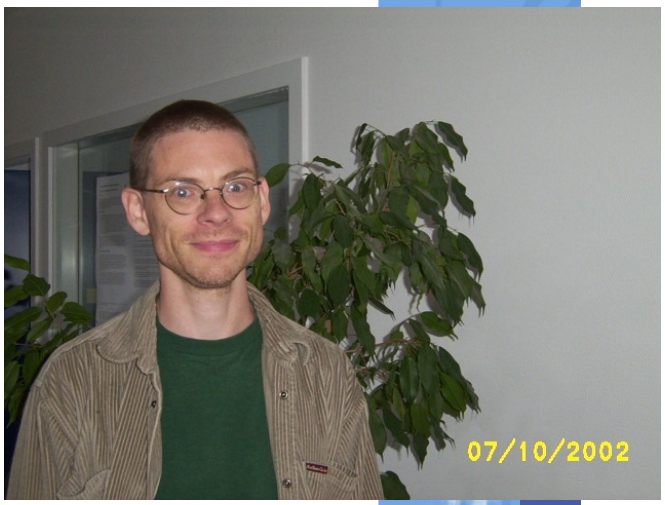
UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

Fredrik Jørgensen (UiO):
Automatisk gjenkjenning av ytringsgrenser i talespråk

- Talespråket er en utfordring som må løses for at man siden skal kunne analysere talespråkets grammatikk automatisk.
- Skriftspråk er tydelig avgrenset ved tegnsetting,
- Transkribert talespråk er ikke inndelt på noen enhetlig måte, så her må det nytenkning til.
- Viktig å utvikle automatiske metoder for å finne ytringsgrenser i NoTa-korpuset,
- Noen foreløpige tall viser hvor vellykket det har blitt.

Peter Juel Henrichsen (KU):
NoTa – nu med lydskrift

- Å lage norsk lydskrift av norsk ortografisk transkripsjon på, ved å gå via dansk!
- Systemet NoTaFon
- Utnytter
  - at NoTa-korpuset er grammatisk tagget med ordklasser og bøying
  - noen enkle ordlister over de vanligste ordene som er mest forskjellige på dansk og norsk.
- Får automatisk laget norske tekst om til dansk tekst
- I dansk kan han så bruke et system som finnes fra før for å gå fra ortografi til lydskrift.
- Setter opp noen enkle regler for hvordan dansk og norsk lyd skiller seg fra hverandre (bløte og harde konsonanter, for eksempel),
- Får et automatiske system fra dansk lydskrift til norsk lydskrift!



Åshild Søfteland og Anders Nøklestad (UiO):

Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger

- Talespråk er fullt av gjentakelser, avbrudd og pauser
- Derfor vanskelig å bruke eksisterende analysemetoder, f.eks. Tagger
- Alternativ: Lage en statistisk tagger for talespråk:
  - Begynn med å tagge en del av teksten manuelt; dette blir det første treningskorpuset
  - Tren taggeren på treningskorpuset
  - Kjør taggeren på en ny del av teksten
  - Rett opp feila som taggeren har gjort
  - Legg den oppretta teksten til treningskorpuset
  - Gå til punkt 2
- Metode: Treetagger. Resultat: 96,9 % korrekt.

UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

Victoria Rosén (UiB):

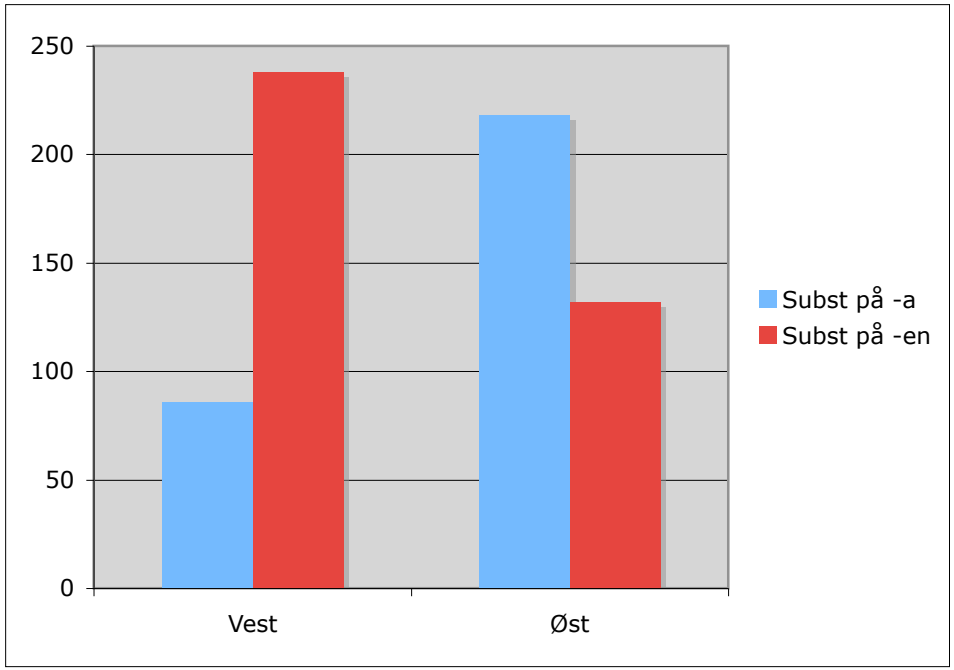
Mot en trebank for talespråk

- Diskuterer viktigheten av å ha en samling av syntaktisk annoterte setninger
  - Viktig for språkteknologisk utivkling
  - Viktig for språkteknologisk evaluering
- Vanskelig med syntaktisk annotering av talespråk
- En løsning: å foreta manuell for-redigering for senere automatisk analyse

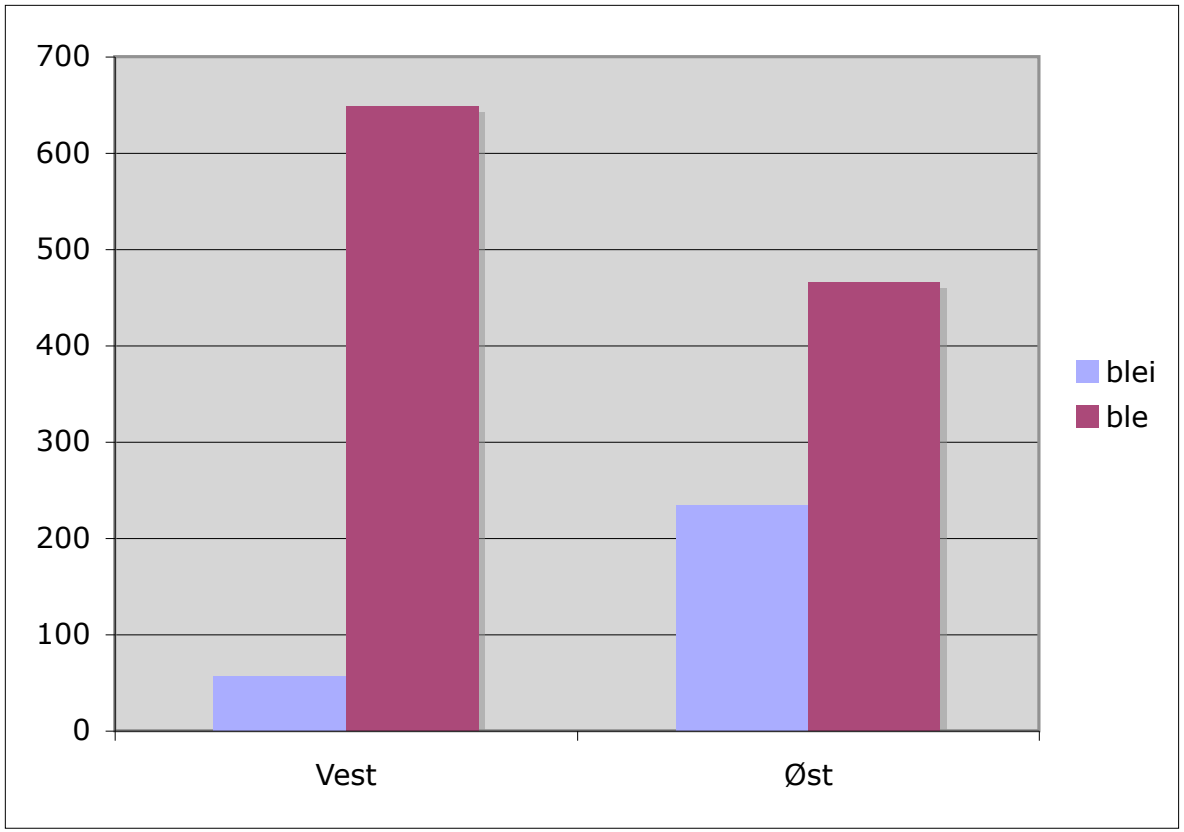
Janne Bondi Johannessen:
Oslospråket i tall

UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

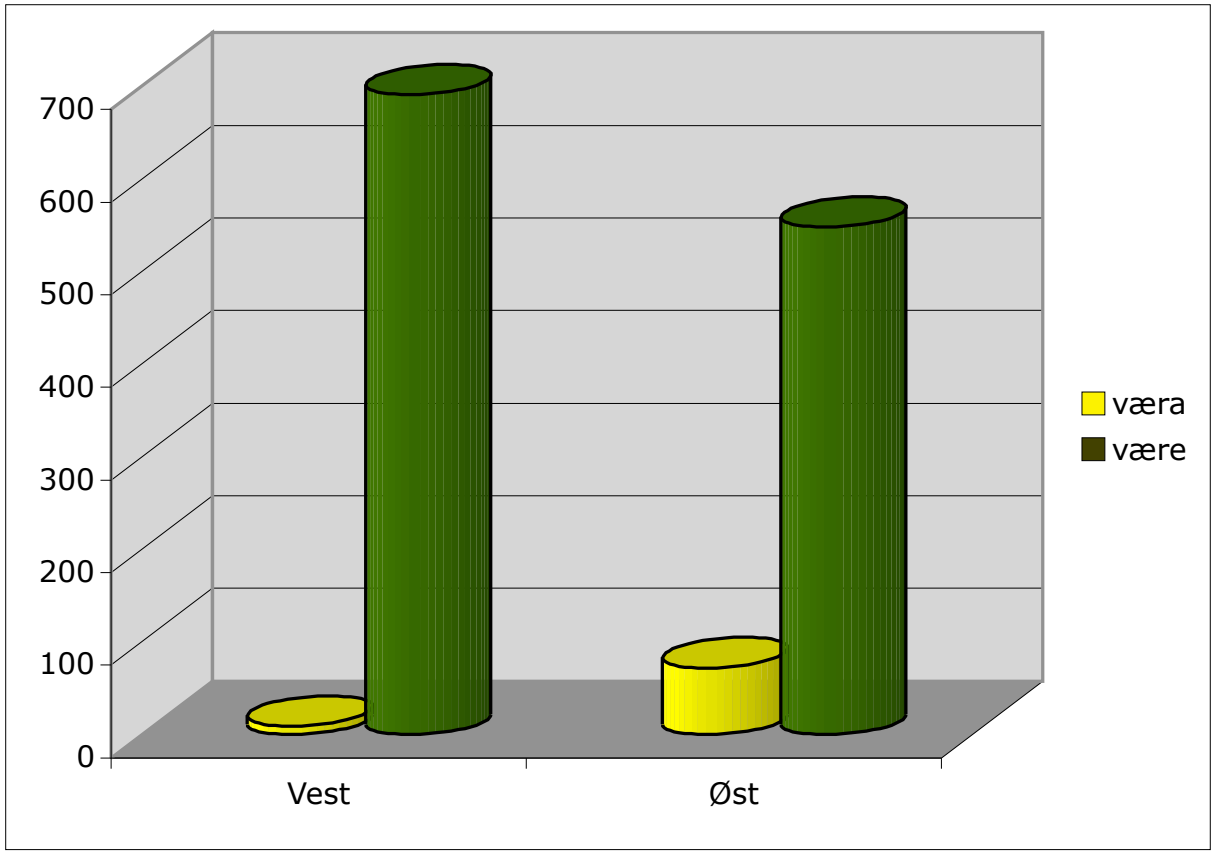
Øst og vest:
Elva eller *elven, avisa* eller *avisen, gata* eller *gaten, boka* eller *boken, tida* eller *tiden, sola* eller *solen?*

- 

Øst og vest: Diftonger – *blei* eller *ble*?

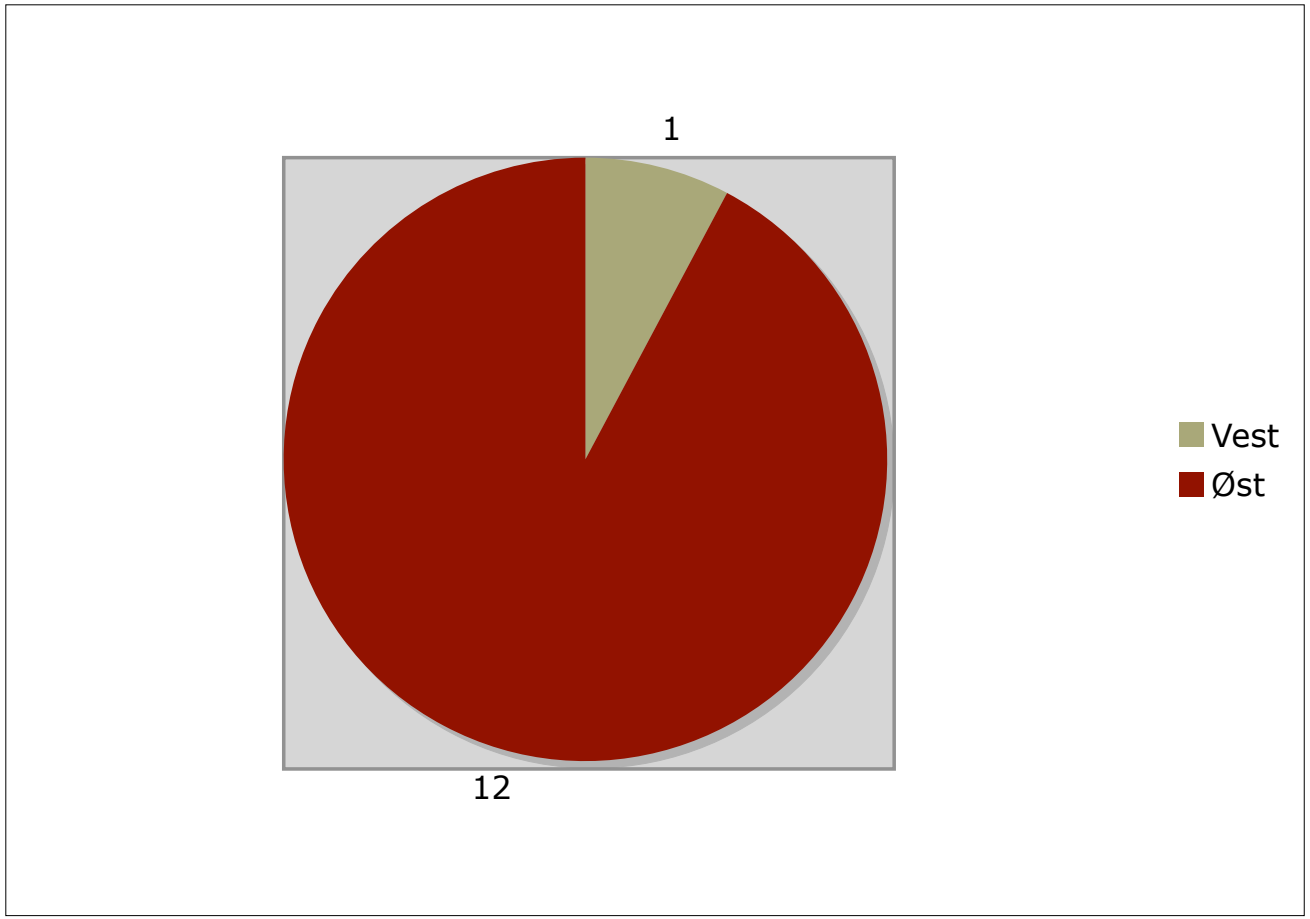


Øst og vest:
Kløyvd infinitiv – *Væra* eller *være*?
(antall forekomster)



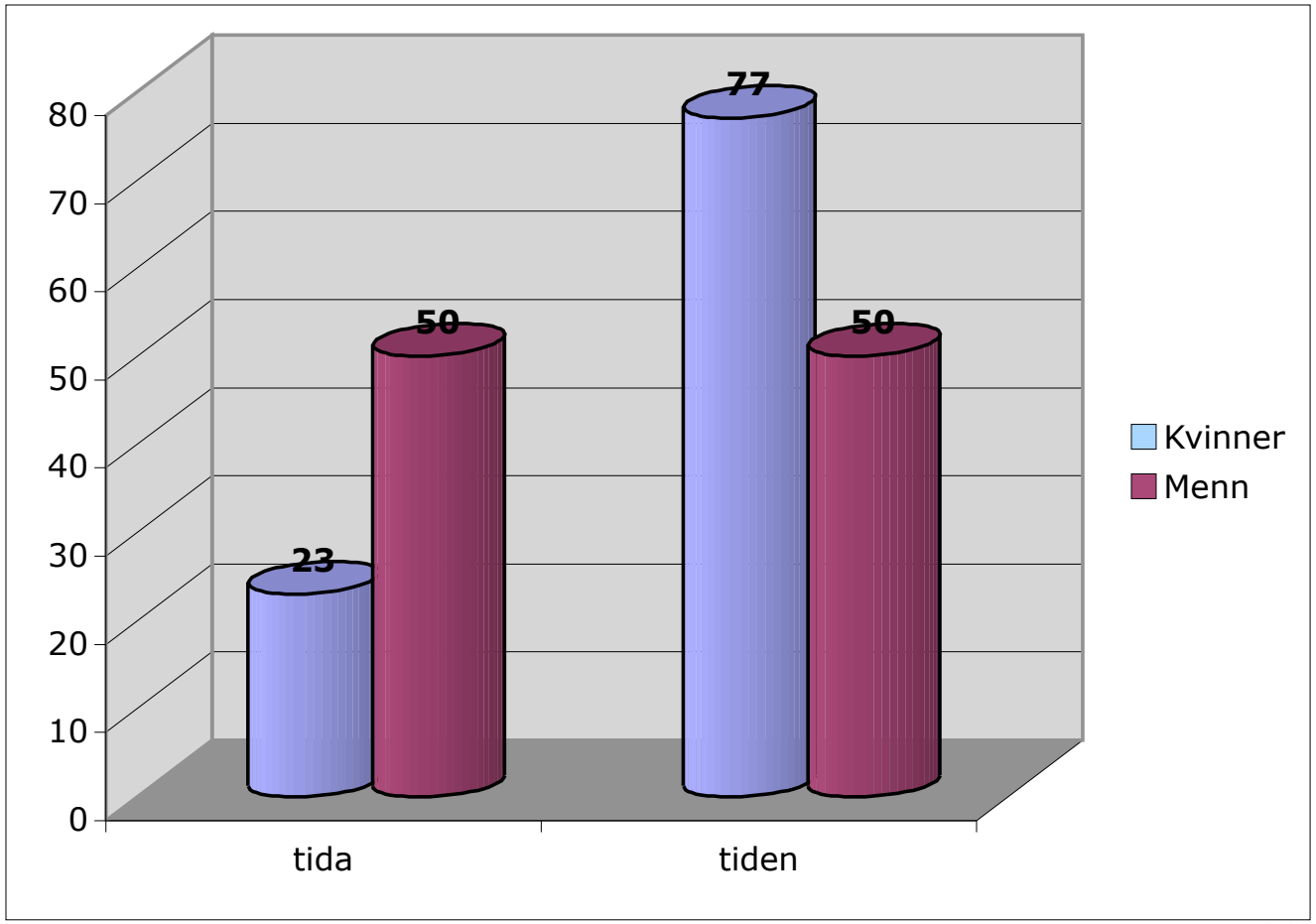
Kløyvd infinitiv i prosent: *væra*

i prosent av alle ganger verbet brukes i begge former i øst og vest

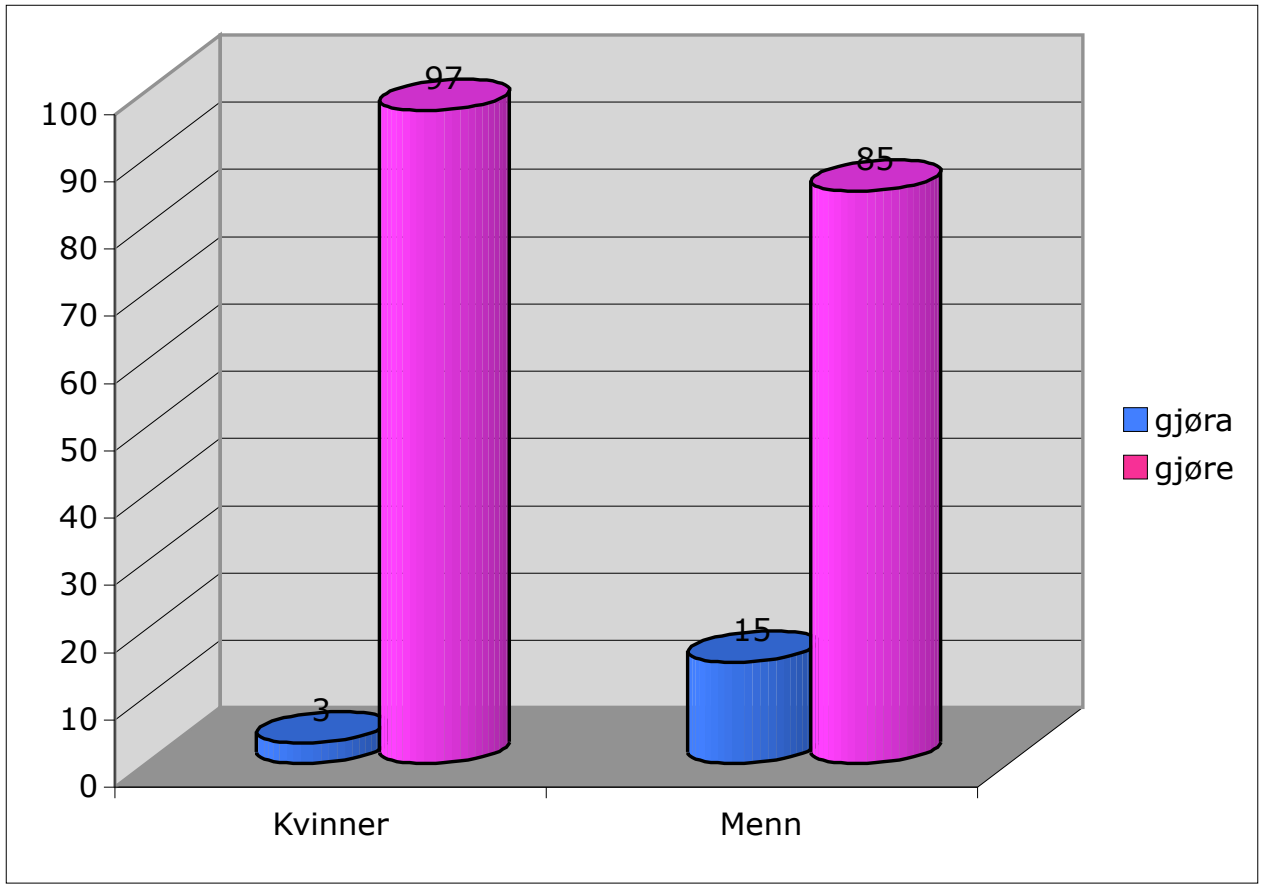


Kjønn:
tida eller *tiden*?

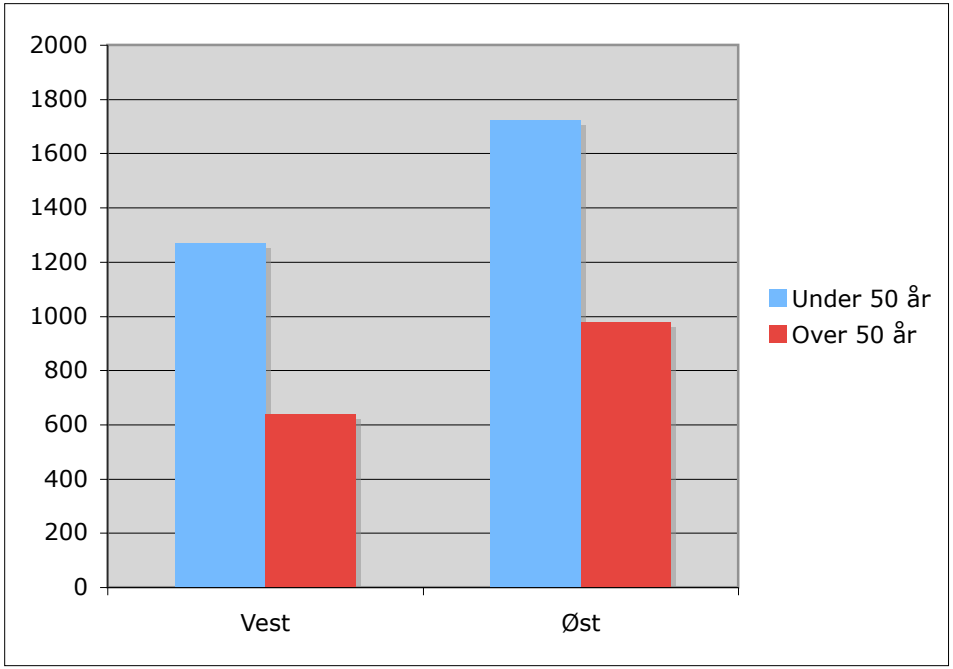
(Prosent, kjønnsvis)



Kjønn:
Kløyvd infinitiv (*gjøra* eller *gjøre*)

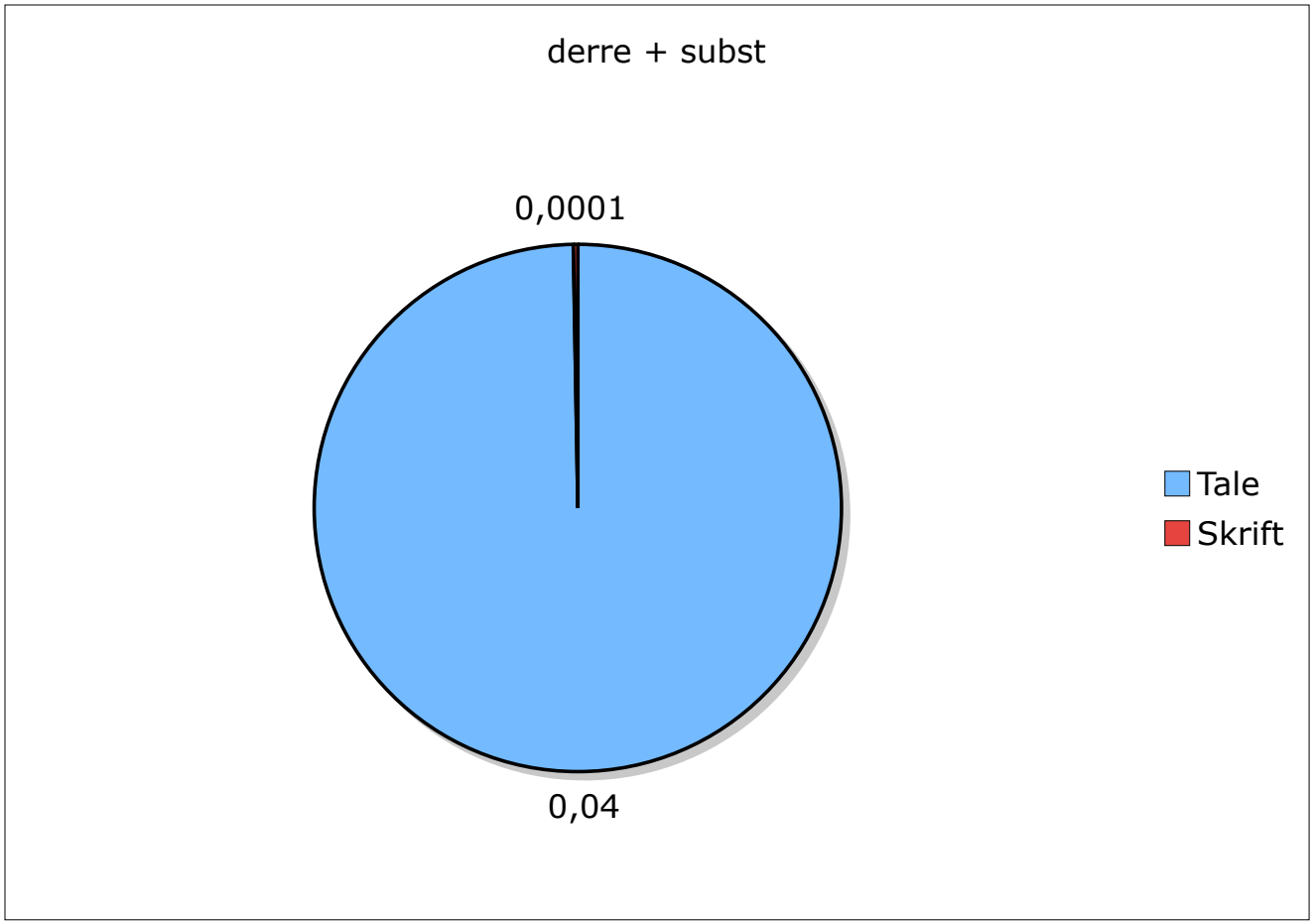


Alder:
Substantiver på -a (*kona, gata*)



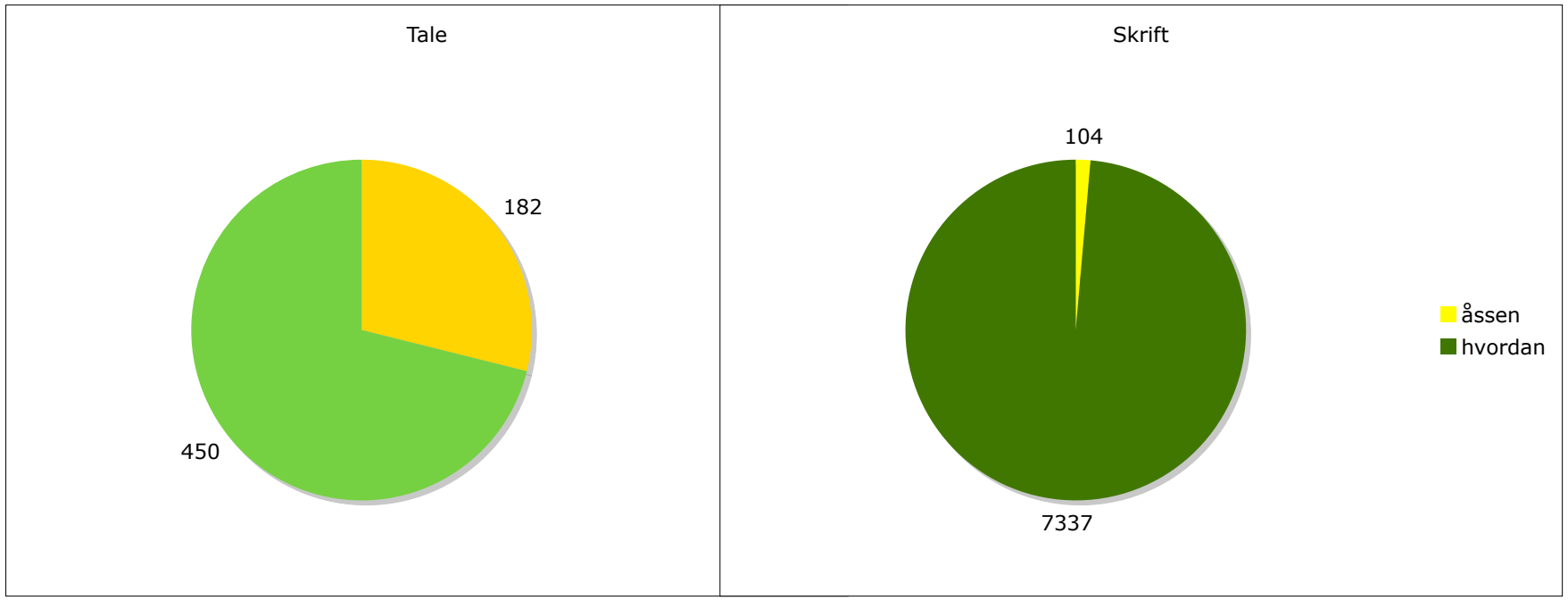
Skrift og tale:
derre + substantiv

(NoTa-korpuset og Oslo-korpuset av taggede, norske tekster)



Tale og skrift:
åssen og *hvordan*

UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET



And from now on: Spoken corpora of some of the Ethiopian languages!