UiO : Det humanistiske fakultet

A multilingual speech resource: The Nordic Dialect Corpus

Janne Bondi Johannessen, Joel Priestley, Anders Nøklestad

Workshop on Advanced Corpus Solutions, PACLIC 24
Tohoku University, Sendai, Japan
4 November 2010

The ScanDiaSyn-project

Two goals:

- Investigate
  - systematically map and study the syntactic variation across the Scandinavian dialect continuum
- Document
  - create a database: Nordic Syntactic Judgements Database
  - create a corpus: Nordic Dialect Corpus
    - Transcribed and tagged speech material linked with audio and video.
    - Web-based with a user friendly interface on the internet.

Iceland
Sverige
Sweden
Suomi
Finland
Norge
Norway
Danmark
Denmark
United Kingdom
Ireland
Беларусь
Belarus
Polska
Poland
Deutschland
Germany
Україна
Ukraine
Österreich
Austria
France
România
Romania
Italia
Italy
España
Spain
Portugal
Ελλάς
Greece
Türkiye
Turkey

Contents of corpus

	Informants	Places	Words
Denmark	75	14	229 909
Faroe Islands	19	5	48 427
Iceland	4	1	10 287
Norway	301	94	1 200 120
Sweden	126	40	299 866
Total	525	154	1 788 609

Collecting the data

- Two informants from the same measure point speak freely for 20-30 minutes
- An informal setting with refreshments
- The informants cannot talk about ”sensitive and confidential information”
- A list of topics is presented to the informant
- Video-recorded
- Transcribed later

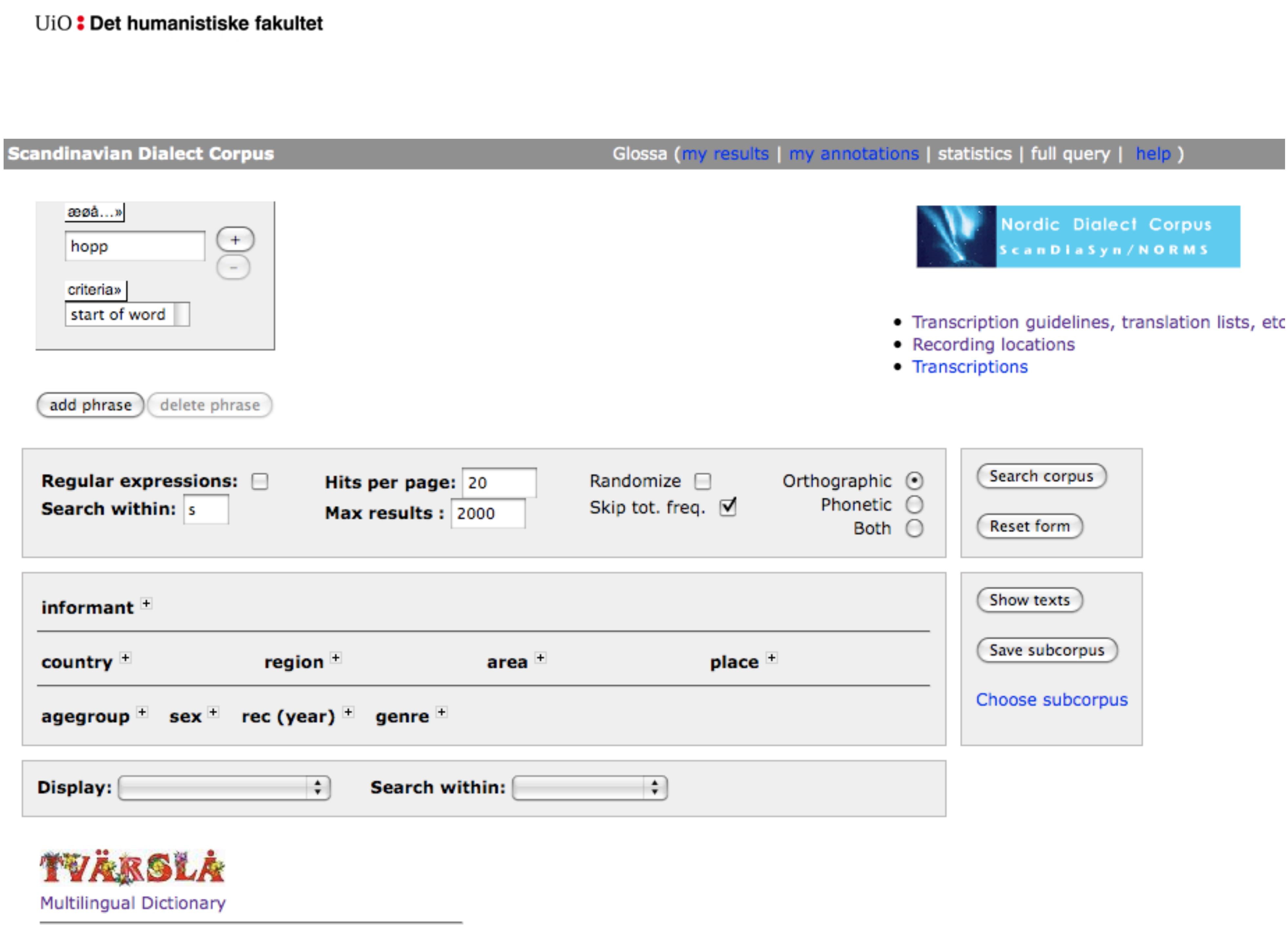
Challenges for a user-friendly system

- five different standard orthographies (Danish, Faroese, Icelandic, Norwegian and Swedish)
- transcribed speech
- some recordings have a double set of transcriptions – orthographic and phonetic
- transcriptions should be linked to audio and video
- the corpus should be tagged, needing five spoken language taggers, but different tagsets
- if possible, the same tags should refer to the same entities
- informant metadata should be filters in search (age, sex etc)

- different levels of geographical belonging should be specifiable (country, area, place)
- all text from all languages should be searchable at the same time
- search results should be possible to handle in a number of different ways, including exporting of different formats
- users want the search results to come with English translation
- the users want maps to see where informants are from

How to search the corpus for all the five languages?

- The five languages are closely related
  - desirable to get all results together
- Still, the orthographies are so different that most people do not know those of the other languages
- Wish from users:
  - a multilingual dictionary that translates all searches
- Solution:
  - a button that links to an online, automatically generated multilingual dictionary. Users choose the appropriate words and puts them into the search interface.



UiO : Det humanistiske fakultet

Scandinavian Dialect Corpus

Glossa (my results | my annotations | statistics | full query | help)

æøå…»

hopp

criteria»

start of word

Regular expressions: □

Search within: s

Hits per page: 20

Max results : 2000

Randomize □

Skip tot. freq. ☑

Orthographic ●

Phonetic ○

Both ○

informant +

country + region + area + place +

agegroup + sex + rec (year) + genre +

Display:

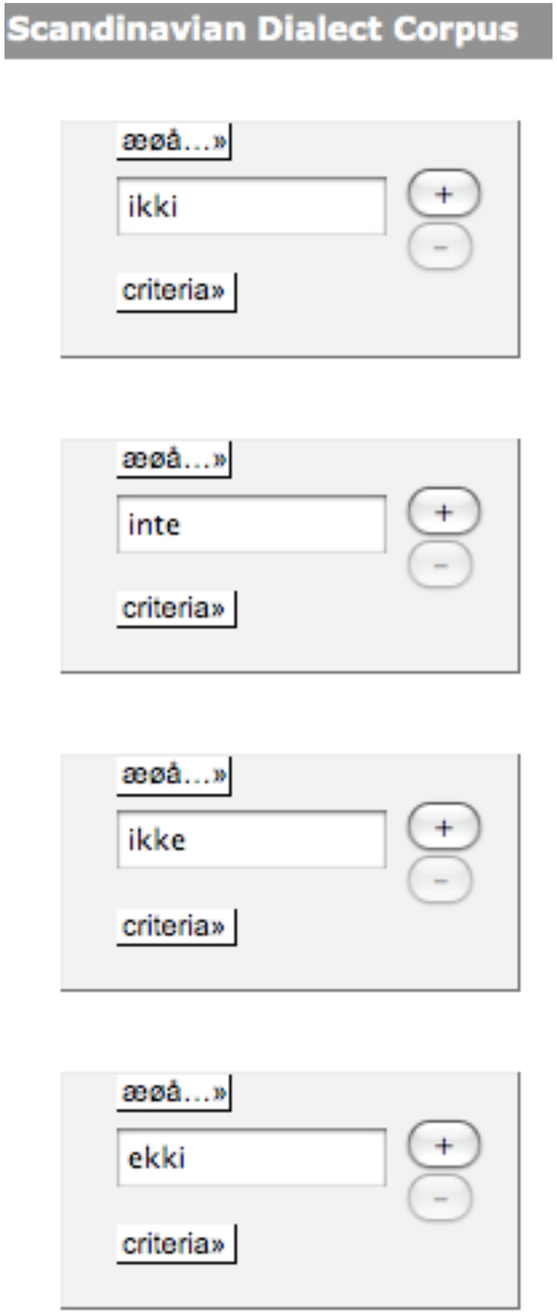
Search within:

Example: the negation word ”not”

- From the multilingual dictionary:
  - *ikke* (Norwegian, Danish)
  - *inte* (Swedish)
  - *ekki* (Icelandic)
  - *ikki* (Faroese)
- Write words in user-friendly boxes.
- Translated to:

(1) **"([((word="ekki" %c))]) | ([((word="inte" %c))]) | ([((word="ikki" %c))]) | ([((word="ikke" %c))]) ;"**

UiO : Det humanistiske fakultet

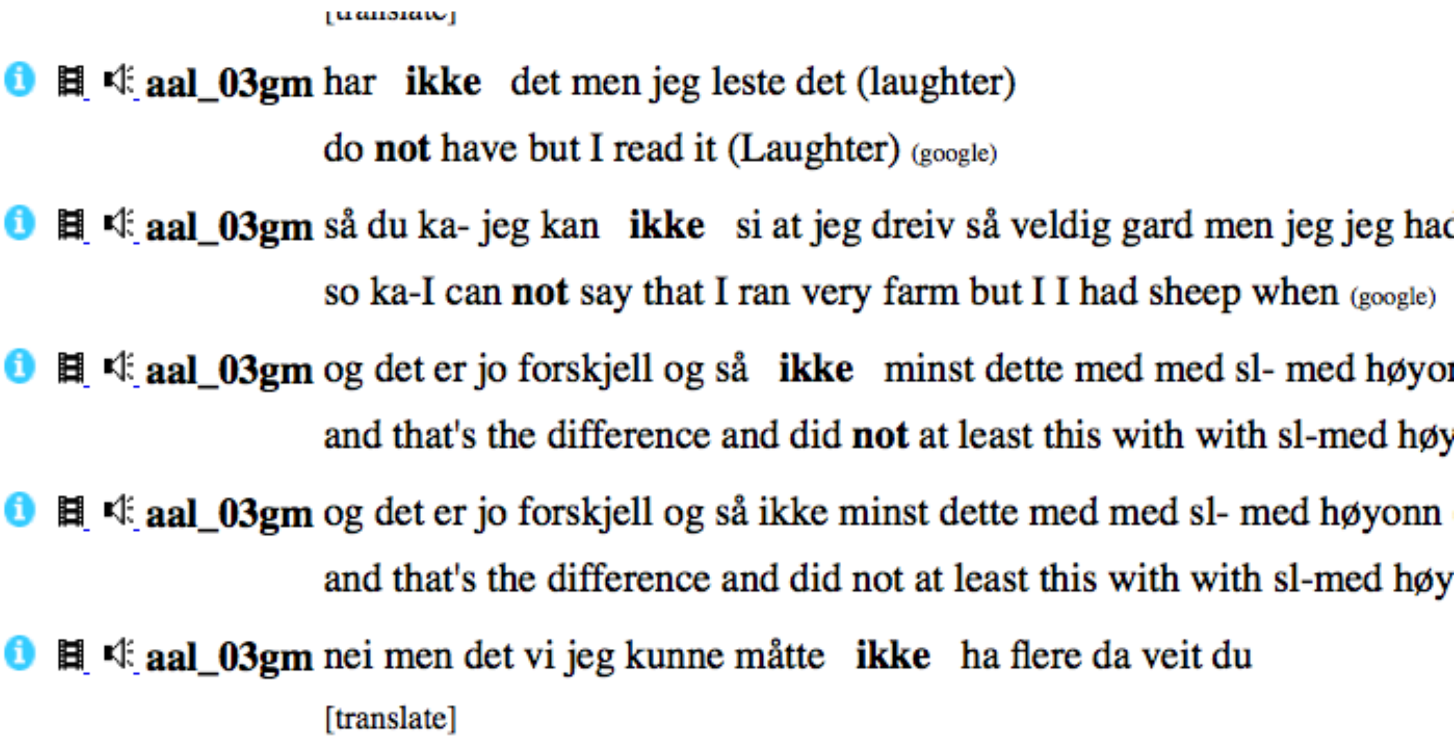

Scandinavian Dialect Corpus

ikki

inte

ikke

ekki


[translate]

aal_03gm har **ikke** det men jeg leste det (laughter)
do **not** have but I read it (Laughter) (google)

aal_03gm så du ka- jeg kan **ikke** si at jeg dreiv så veldig gard men jeg jeg had
so ka-I can **not** say that I ran very farm but I I had sheep when (google)

aal_03gm og det er jo forskjell og så **ikke** minst dette med med sl- med høyo
and that's the difference and did **not** at least this with with sl-med høy

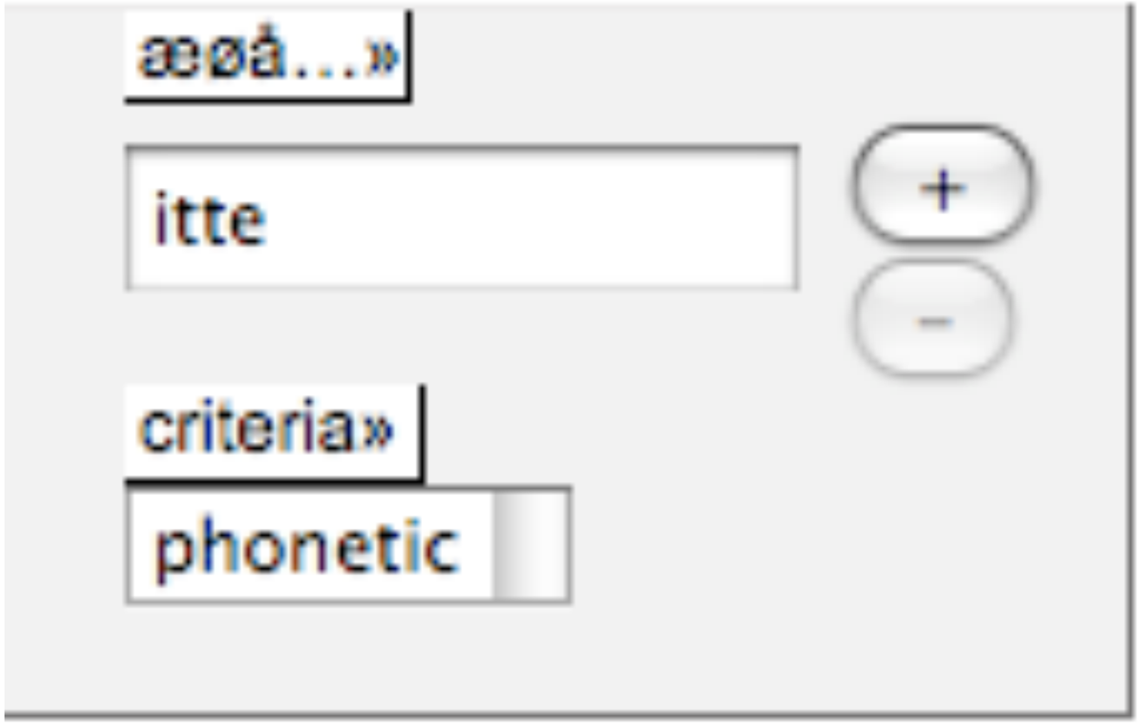
aal_03gm og det er jo forskjell og så ikke minst dette med med sl- med høyonn
and that's the difference and did not at least this with with sl-med høy

aal_03gm nei men det vi jeg kunne måtte **ikke** ha flere da veit du
[translate]

Two transcriptions

- Two transcriptions (Norwegian and some Swedish)
  - Orthographic
  - Phonetic
- First: Speech was transcribed phonetically
- Then: Orthographic transcription was created from the phonetic transcription by a semi-automatic transliterator.
- Totally aligned

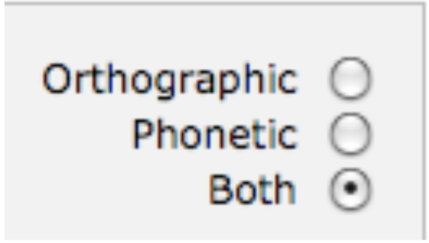
Corpus can be searched phonetically or orthographically



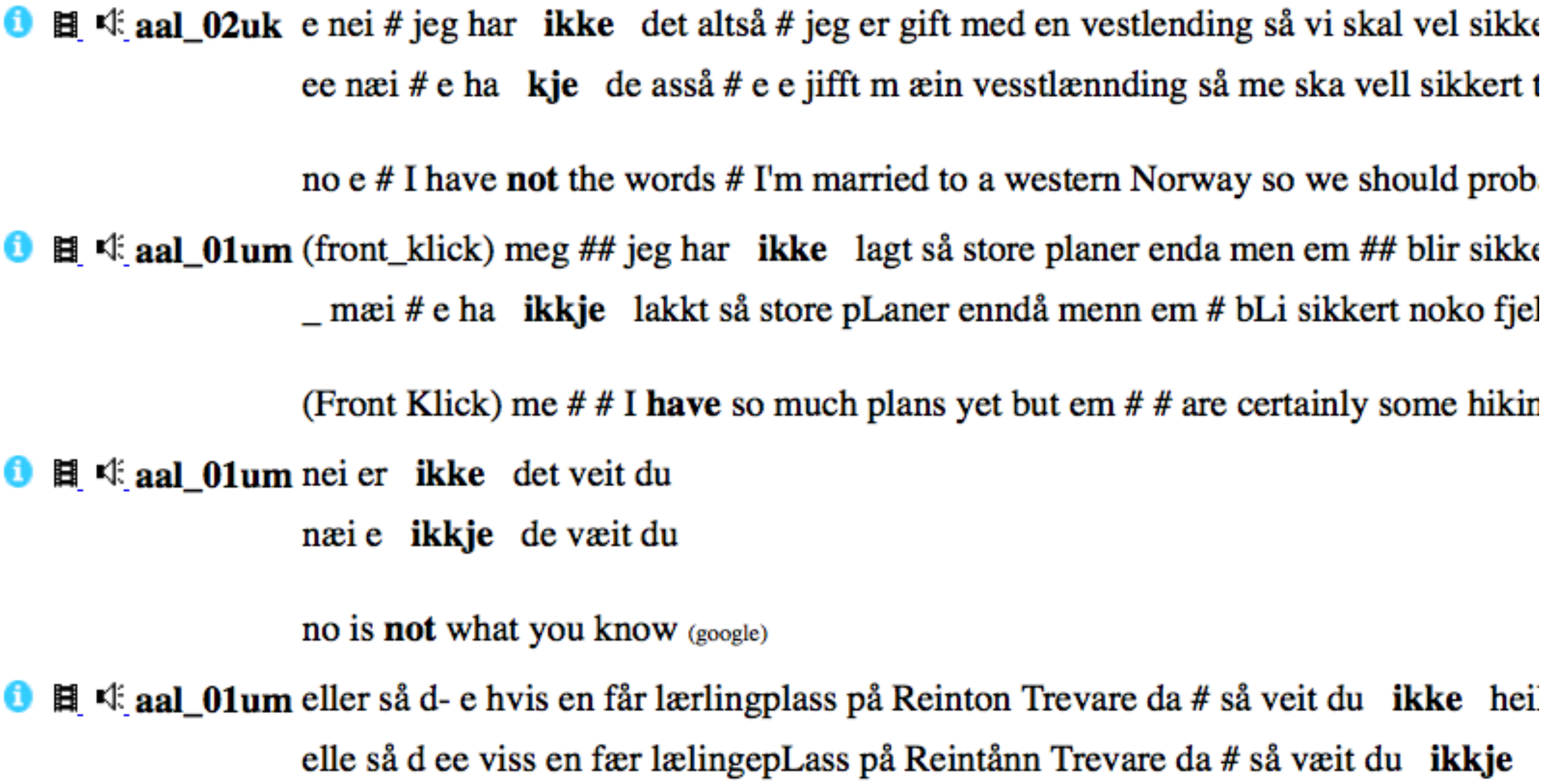
itte

phonetic

Results can be viewed phonetically, orthographically or both



Orthographic ○
Phonetic ○
Both ●



aal_02uk e nei # jeg har **ikke** det altså # jeg er gift med en vestlending så vi skal vel sikke
ee næi # e ha **kje** de asså # e e jifft m æin vesstlænnding så me ska vell sikkert t

no e # I have **not** the words # I'm married to a western Norway so we should prob

aal_01um (front_klick) meg ## jeg har **ikke** lagt så store planer enda men em ## blir sikke
_ mæi # e ha **ikkje** lakkt så store pLaner enndå menn em # bLi sikkert noko fjel

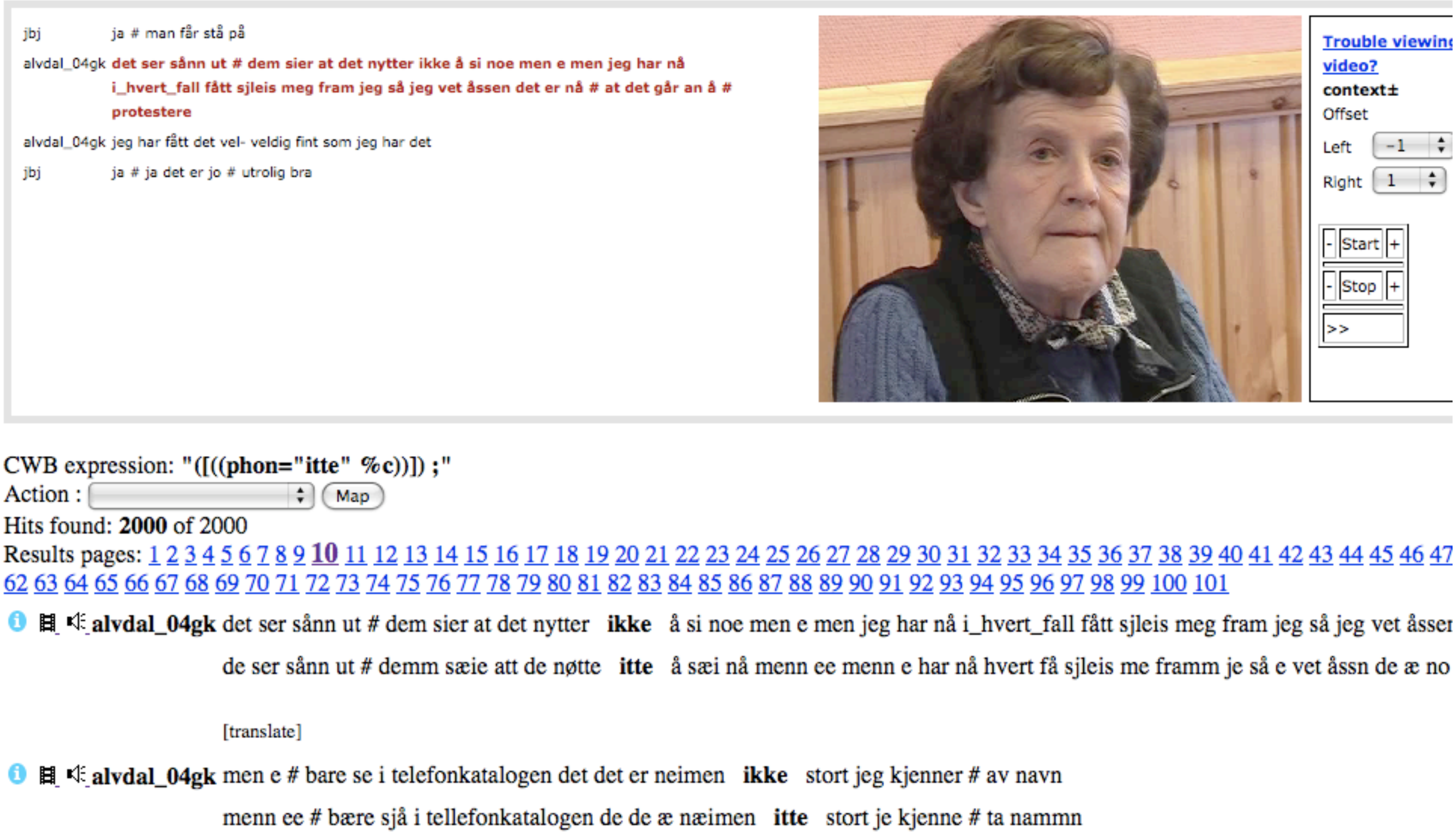
(Front Klick) me # # I **have** so much plans yet but em # # are certainly some hikin

aal_01um nei er **ikke** det veit du
næi e **ikkje** de væit du

no is **not** what you know (google)

aal_01um eller så d- e hvis en får lærlingplass på Reinton Trevare da # så veit du **ikke** hei
elle så d ee viss en fær lælingepLass på Reintånn Trevare da # så væit du **ikkje**

Links from hits to audio and video



jbj ja # man får stå på

alvdal_04gk **det ser sånn ut # dem sier at det nytter ikke å si noe men e men jeg har nå i_hvert_fall fått sjleis meg fram jeg så jeg vet åssen det er nå # at det går an å # protestere**

alvdal_04gk jeg har fått det vel- veldig fint som jeg har det

jbj ja # ja det er jo # utrolig bra

CWB expression: "([((phon="itte" %c))]) ;"

Action :

Hits found: **2000** of 2000

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 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101

alvdal_04gk det ser sånn ut # dem sier at det nytter **ikke** å si noe men e men jeg har nå i_hvert_fall fått sjleis meg fram jeg så jeg vet åsser

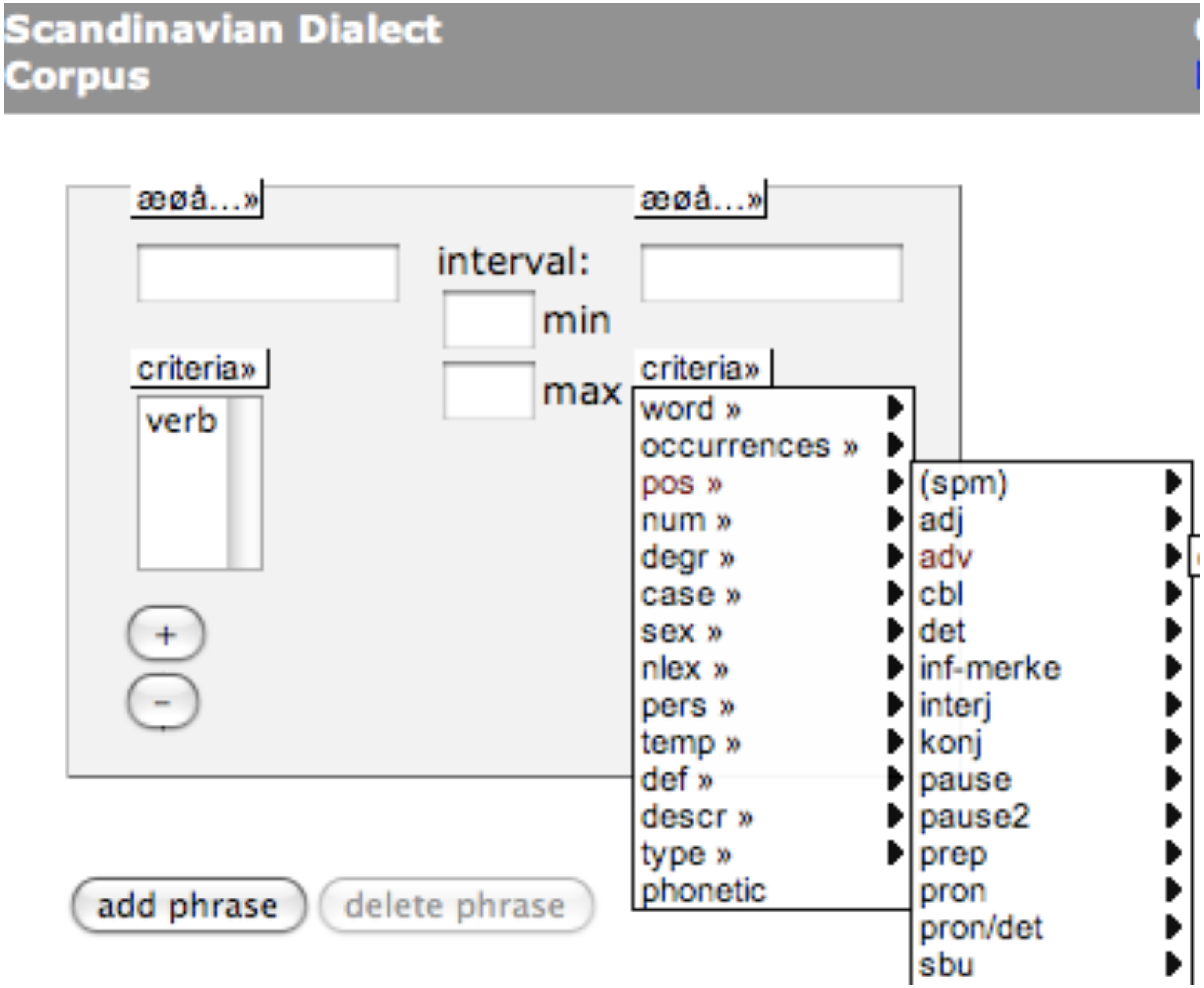
de ser sånn ut # demm sæie att de nøtte **itte** å sæi nå menn ee menn e har nå hvert få sjleis me framm je så e vet åssn de æ no

[translate]

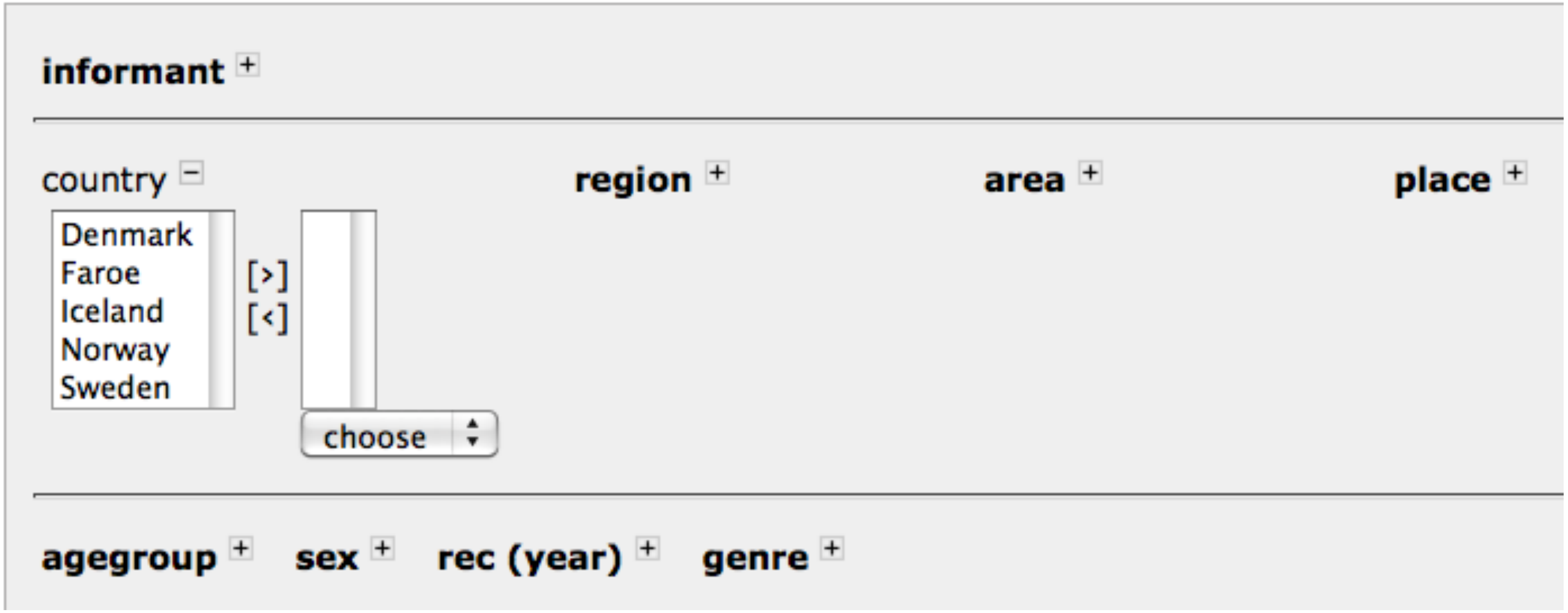
alvdal_04gk men e # bare se i telefonkatalogen det det er neimen **ikke** stort jeg kjenner # av navn

menn ee # bære sjå i tellefonkatalogen de de æ næimen **itte** stort je kjenne # ta nammn

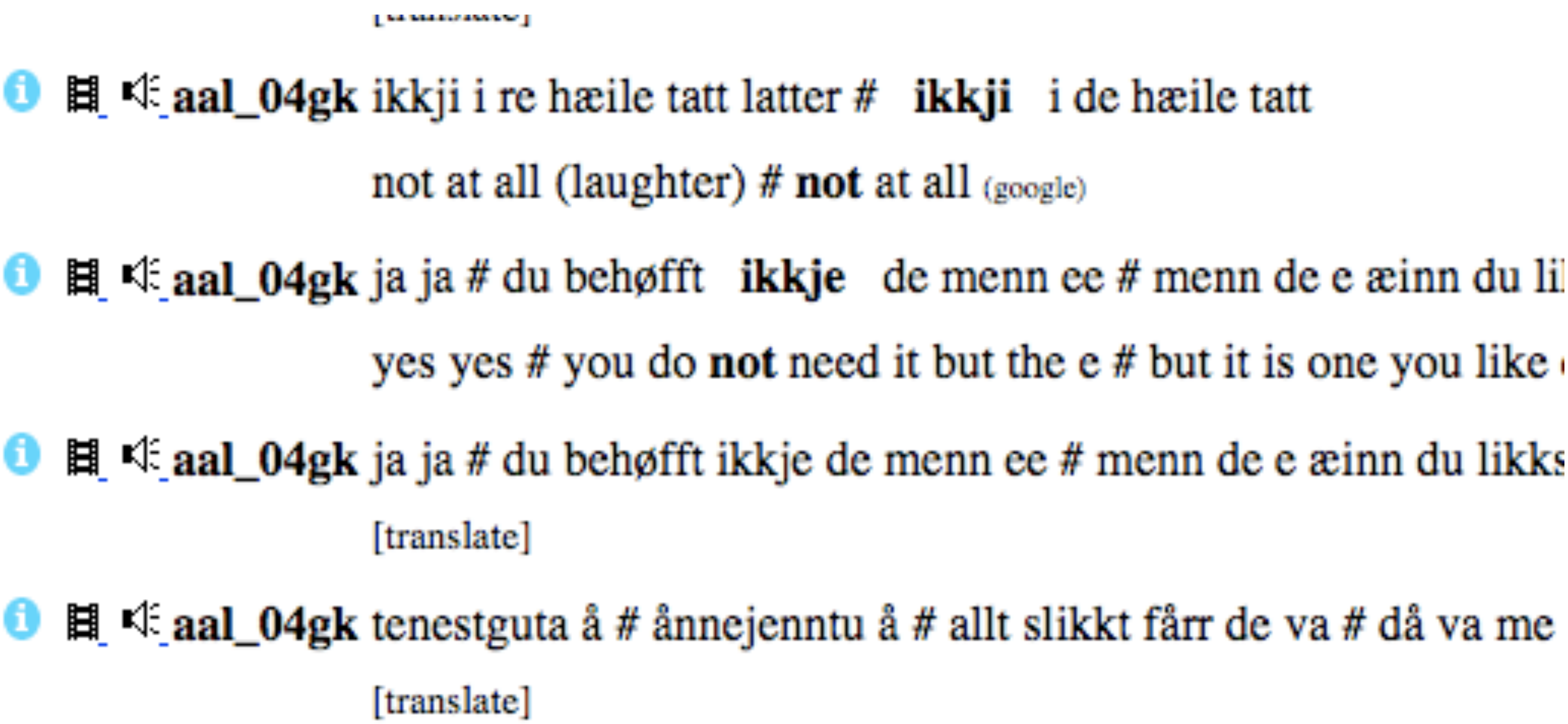
All languages are being tagged.
Tags should be standardised for all.



Metadata filter



All dialects can be translated online to English with Google Translate



[translate]

aal_04gk ikkji i re hæile tatt latter #   **ikkji**   i de hæile tatt

not at all (laughter) # **not** at all (google)

aal_04gk ja ja # du behøfft   **ikkje**   de menn ee # menn de e æinn du li

yes yes # you do **not** need it but the e # but it is one you like

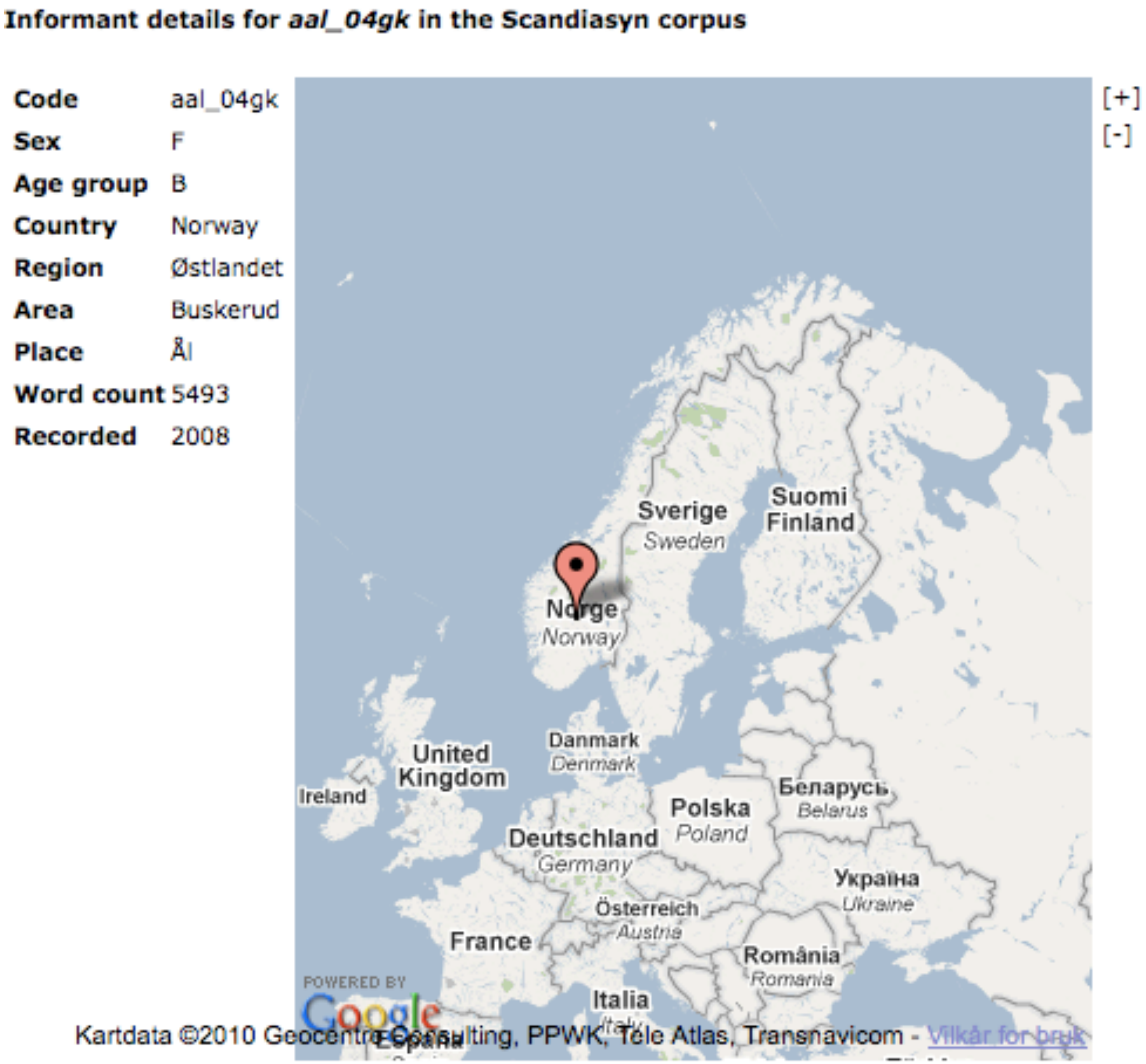
aal_04gk ja ja # du behøfft ikkje de menn ee # menn de e æinn du likks

[translate]

aal_04gk tenestguta å # ånnejenntu å # allt slikkt fårr de va # då va me

[translate]

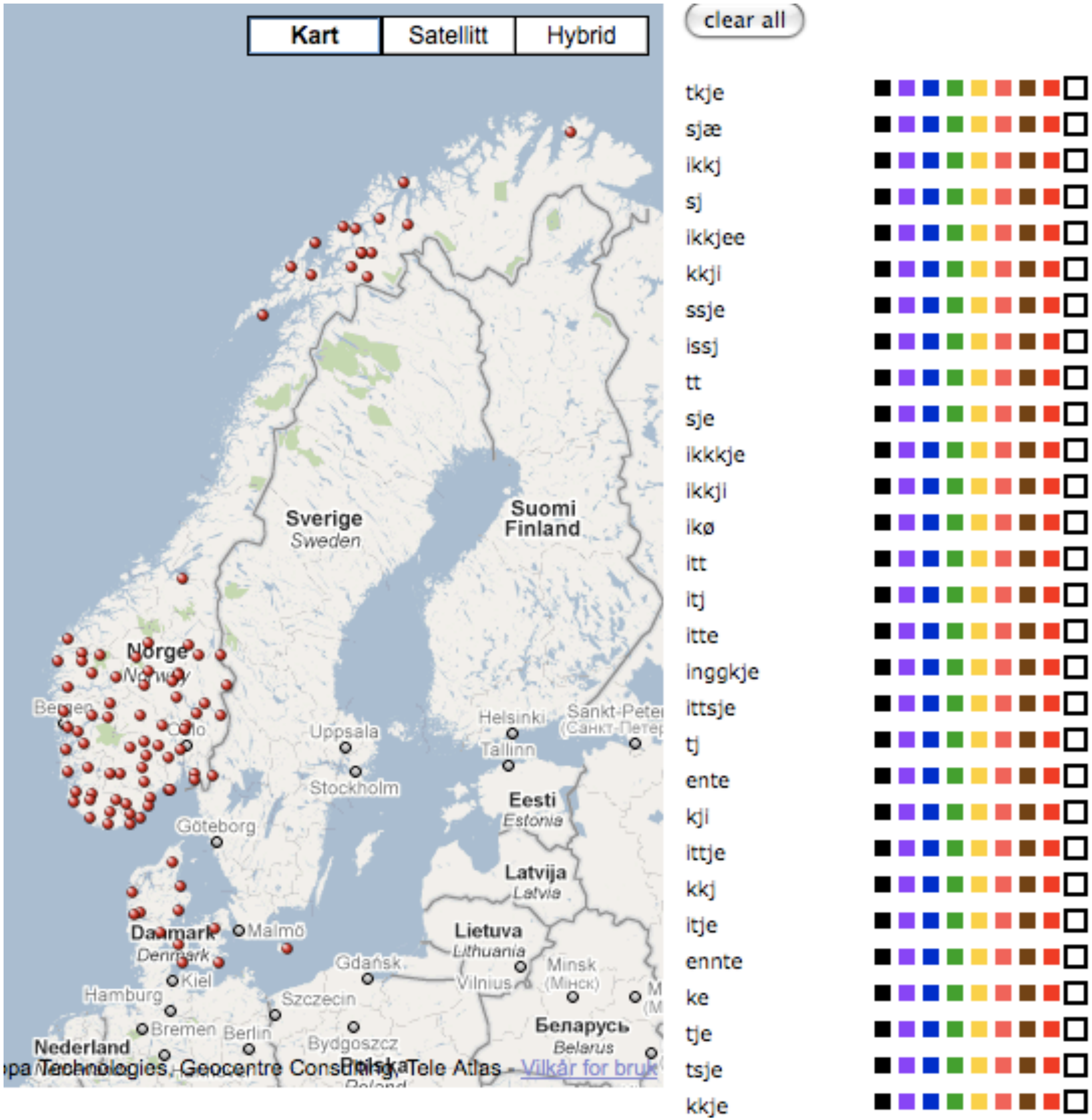
Maps from Google Maps Map view accompanies information-button



Informant details for *aal_04gk* in the Scandiasyn corpus

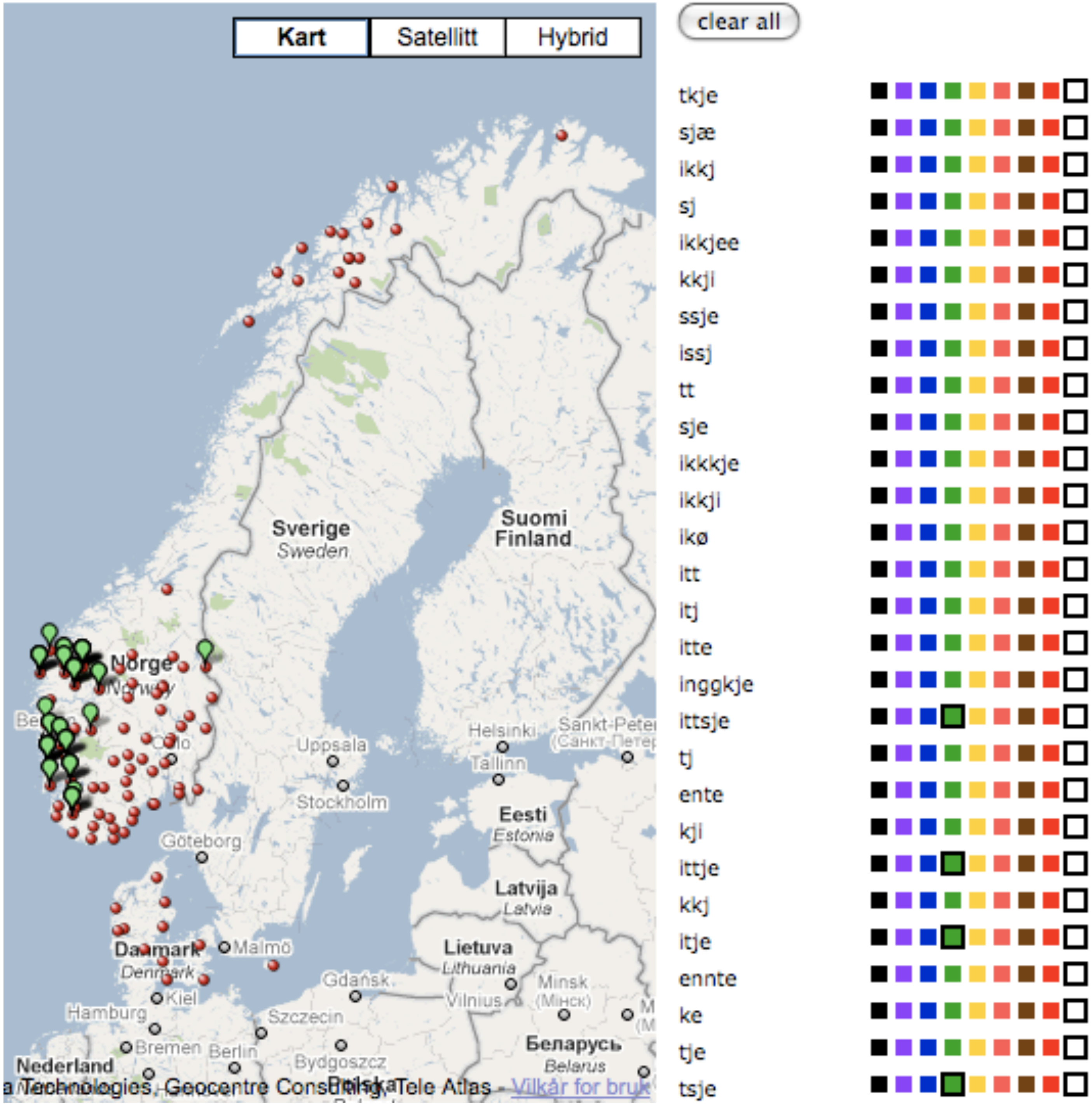
Code	aal_04gk
Sex	F
Age group	B
Country	Norway
Region	Østlandet
Area	Buskerud
Place	Ål
Word count	5493
Recorded	2008

Kartdata ©2010 Geocentre Consulting, PPWK, Tele Atlas, Transnavicom - Vilkår for bruk



tkje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
sjæ ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ikkj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
sj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ikkjee ⬛🟪🟦🟩🟨🟥🟫🟥⬜
kkji ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ssje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
issj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
tt ⬛🟪🟦🟩🟨🟥🟫🟥⬜
sje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ikkkje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ikkji ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ikø ⬛🟪🟦🟩🟨🟥🟫🟥⬜
itt ⬛🟪🟦🟩🟨🟥🟫🟥⬜
itj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
itte ⬛🟪🟦🟩🟨🟥🟫🟥⬜
inggkje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ittsje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
tj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ente ⬛🟪🟦🟩🟨🟥🟫🟥⬜
kji ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ittje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
kkj ⬛🟪🟦🟩🟨🟥🟫🟥⬜
itje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ennte ⬛🟪🟦🟩🟨🟥🟫🟥⬜
ke ⬛🟪🟦🟩🟨🟥🟫🟥⬜
tje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
tsje ⬛🟪🟦🟩🟨🟥🟫🟥⬜
kkje ⬛🟪🟦🟩🟨🟥🟫🟥⬜

Maps from Google depict where the hits have been found



tkje	■ ■ ■ ■ ■ ■ ■ ■ □
sjæ	■ ■ ■ ■ ■ ■ ■ ■ □
ikkj	■ ■ ■ ■ ■ ■ ■ ■ □
sj	■ ■ ■ ■ ■ ■ ■ ■ □
ikkjee	■ ■ ■ ■ ■ ■ ■ ■ □
kkji	■ ■ ■ ■ ■ ■ ■ ■ □
ssje	■ ■ ■ ■ ■ ■ ■ ■ □
issj	■ ■ ■ ■ ■ ■ ■ ■ □
tt	■ ■ ■ ■ ■ ■ ■ ■ □
sje	■ ■ ■ ■ ■ ■ ■ ■ □
ikkkje	■ ■ ■ ■ ■ ■ ■ ■ □
ikkji	■ ■ ■ ■ ■ ■ ■ ■ □
ikø	■ ■ ■ ■ ■ ■ ■ ■ □
itt	■ ■ ■ ■ ■ ■ ■ ■ □
itj	■ ■ ■ ■ ■ ■ ■ ■ □
itte	■ ■ ■ ■ ■ ■ ■ ■ □
inggkje	■ ■ ■ ■ ■ ■ ■ ■ □
ittsje	■ ■ ■ [✓ ■] ■ ■ ■ ■ □
tj	■ ■ ■ ■ ■ ■ ■ ■ □
ente	■ ■ ■ ■ ■ ■ ■ ■ □
kji	■ ■ ■ ■ ■ ■ ■ ■ □
ittje	■ ■ ■ [✓ ■] ■ ■ ■ ■ □
kkj	■ ■ ■ ■ ■ ■ ■ ■ □
itje	■ ■ ■ [✓ ■] ■ ■ ■ ■ □
ennte	■ ■ ■ ■ ■ ■ ■ ■ □
ke	■ ■ ■ ■ ■ ■ ■ ■ □
tje	■ ■ ■ ■ ■ ■ ■ ■ □
tsje	■ ■ ■ [✓ ■] ■ ■ ■ ■ □

Maps make it possible to show where specific phonetic realisations have been found.

What is the system used for the corpus?

- Glossa
  - Developed at the Text Laboratory, UiO
  - Front end: menus and boxes
  - Middle end: MySQL database for metadata
  - Back end: Corpus Work Bench (CWB, CQP)
  - Is being modularised at the moment for freer combinations with other back ends.
  - Freely downloadable
  - Assistance provided

- http://tekstlab.uio.no/glossa//html/GLOSSA_manual.html

- http://www.hf.uio.no/iln/om/organisasjon/tekstlab/