Research Infrastructure for Linguistic Variation Studies

RILiVS

Øystein A. Vangsnes

CASTL, University of Tromsø

RILiVS

… is a series of exploratory workshops funded by NOS-HS.

… is an initiative among four Nordic research networks/ projects which all build or aim to build infrastructures relating to language variation.

Main objectives:

- *consolidate the resources developed in the individual projects*
- *establish common and realistic research goals in the field of linguistic variation*
- *develop common standards and tools to enhance the quality of linguistic variation research*

The RILiVS Basis

The Medieval Nordic Text Archive (Menota):

17 Nordic partners/institutions

Scandinavian Dialect Syntax (ScanDiaSyn):

10 Nordic research groups (plus EU extensions)

Swedish Dialect Database (SweDat):

Continuation of SweDia 2000; collaborative Swedish project

Saami Documentation and Revitalization Network (SaamiDocNet):

27 groups across 8 countries

Three meetings and the outcome

1. Consolidation (Gothenburg, November 2008)
2. Outreach (Oslo, September 2009)
3. Concretization (Reykjavík?, soon…)

- What are viable and feasible collaborative projects?
- Nordic? European?
- RILiVS should lead to concrete project applications.
- FP7-INFRASTRUCTURES-2010-2:
  24 November 2009, at 17.00.00, Brussels local time.

The linguist's perspective

Important!

Language vs. language variation

Are there any special infrastructure needs when it comes to dealing with and studying linguistic variation?

Macro- vs. microvariation?

Topics

From the CfP for the Oslo meeting:

– *Corpus search interfaces and results handling*

– *Language maps/cartography*

– *Annotation (for grammar as well as for metadata)*

– *Standardization (fonts)*

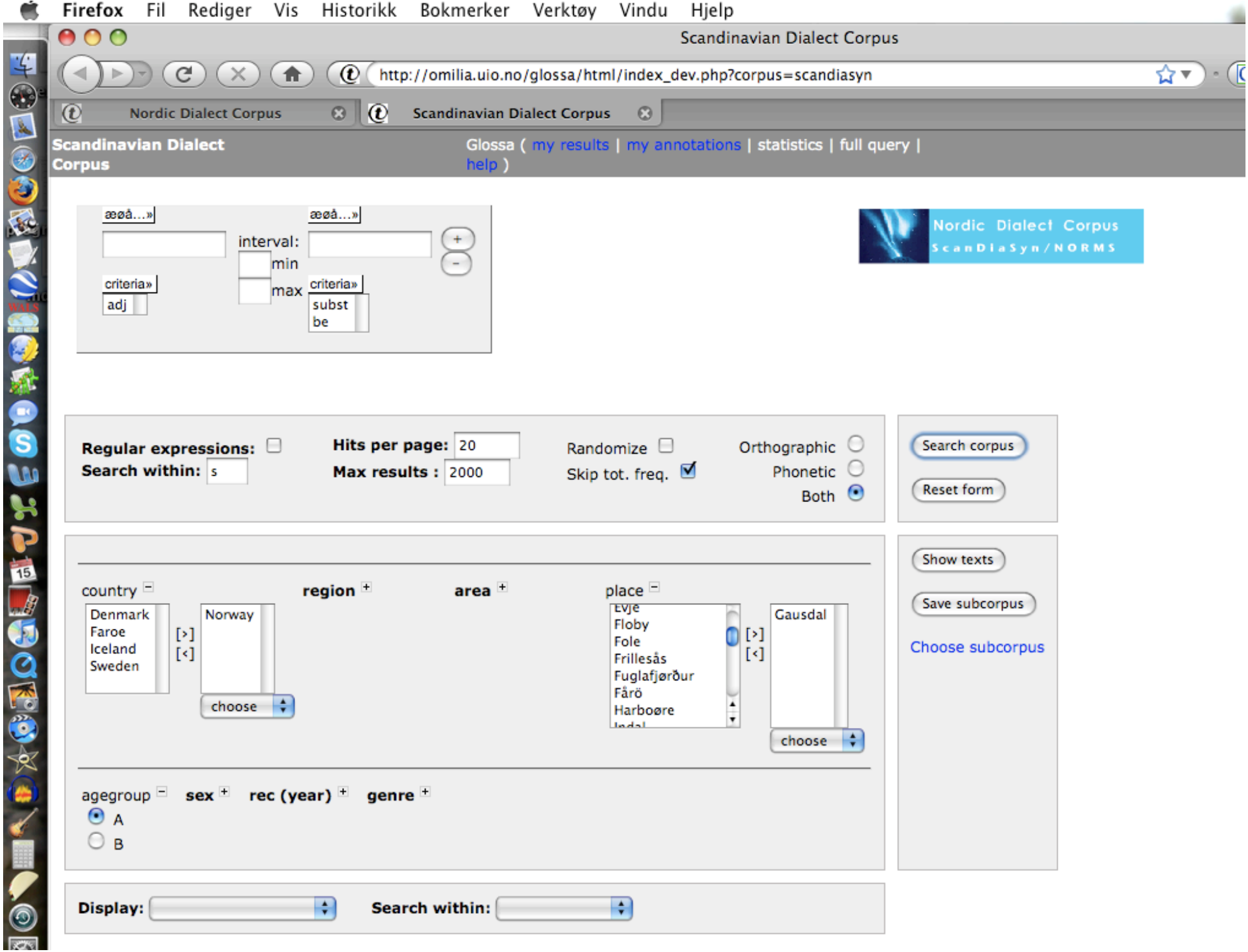
The ScanDiaSyn infrastructure

- The Nordic Dialect Corpus:

  Transcribed and tagged free speech dialect recordings (interviews and conversation) from >200 measure points across the Nordic countries, searchable online and linked up with audio/video.

- The Nordic Syntactic Judgments Database

  Database of results from questionnaire based inquiries on syntactic and morphosyntactic structures.


Firefox Fil Rediger Vis Historikk Bokmerker Verktøy Vindu Hjelp
Scandinavian Dialect Corpus
http://omilia.uio.no/glossa/html/index_dev.php?corpus=scandiasyn
Nordic Dialect Corpus
Scandinavian Dialect Corpus
Scandinavian Dialect Corpus
Glossa (my results | my annotations | statistics | full query | help)
æøå...»
interval:
min
max
criteria»
adj
æøå...»
criteria»
subst
be
Nordic Dialect Corpus
ScanDiaSyn/NORMS
Regular expressions:
Search within:
s
Hits per page:
20
Max results :
2000
Randomize
Skip tot. freq.
Orthographic
Phonetic
Both
Search corpus
Reset form
country
Denmark
Faroe
Iceland
Sweden
[>]
[<]
Norway
choose
region
area
place
Evje
Floby
Fole
Frillesås
Fuglafjørður
Fårö
Harboøre
[>]
[<]
Gausdal
choose
Show texts
Save subcorpus
Choose subcorpus
agegroup
A
B
sex
rec (year)
genre
Display:
Search within:

Informants: 2
scandiasyn:
CWB expression: "([((pos="adj"))][((defn="be") & (pos="subst"))]) ;"
Action :
: **20**
Results pages: 1 2

gausdal_01um det var det # og jeg var på Otta på **første sørvisen** nå i ...
dæ va re # å e va på Otta på **fysste sørrvisen** nå i ...

gausdal_05um ja # ja ja åssen er det dere fyrer med ved må dere holde på å legge innatt i **hele tida** da ?
ja # ja ja åss`n æ re døkk fyre me ve må røkk hæill på å lægge innatt i **hæle tia** ra ?

gausdal_01um nei for vi har m setter opp i anlegg som skal være # nok **åt bygningen** hjemme
næi førr vi har m sætt opp i annlegg såmm ska vera # nåkk **åt byggningen** heme

gausdal_01um og der skal det være **varme golv** der # så støkkte vi ned rør i mjølkerommet vi har i omkledningsrom i fjøset nå
å dær ska re være **varrme gåLLv** der # så støkkte vi ne røyrer i mjøLLkeromme vi ha i ommkLedningsromm i fjøse no

gausdal_05um er ikke **rare ørene** du får utav veit du slik # en slik pelletsovn da vet du
e itte **rare øørn** su fær uta væit du sjlek # en sjlek pelletsåmm da væt du

gausdal_01um ja ja jeg har jo lånt mye penger de **siste årene** nå kan du si
ja ja e ha jo lånt mye pænnger di **sisste åran** nå kænn du si

gausdal_01um og da har dem # funnet at smågris med sviskader # mens i # i vannvarme så går det # **hele tida** i ...
å da ha rømm # funni att smågris me sviskader # menns i # i vassværrme så går de # **hæle tia** i ...

gausdal_05um ja det går der **hele tida** det veit du ## det går der hele tida da vet du
ja de går der **hæle tia** ræ væit du ## de går dær hæle tia ra væt du

gausdal_05um ja det går der hele tida det veit du ## det går der **hele tida** da vet du
ja de går der hæle tia ræ væit du ## de går dær **hæle tia** ra væt du

gausdal_01um på barneskolen da var det n- om vinteren var det ski ## så e ## det var ski og slik **heile tida**

Issues in the ScanDiaSyn corpus

- *Semi-automatic transcription and automatic tagging lead to errors.*

- *The corpus is useless without user proficiency in the languages it contains.*

  (Is Google Translate the savior?)

The Menota corpus

- 17 medieval Nordic texts (approx. 923,000 words).
- Texts displayed on one or more levels: facsimile (very close transcription), diplomatic (less close and with some interpretation), normalised (regularised orthography).
- Pdf-generation offered to avoid font problems.
- Some texts are linked up with a dictionary.

Let's go surfing!

Menota issues

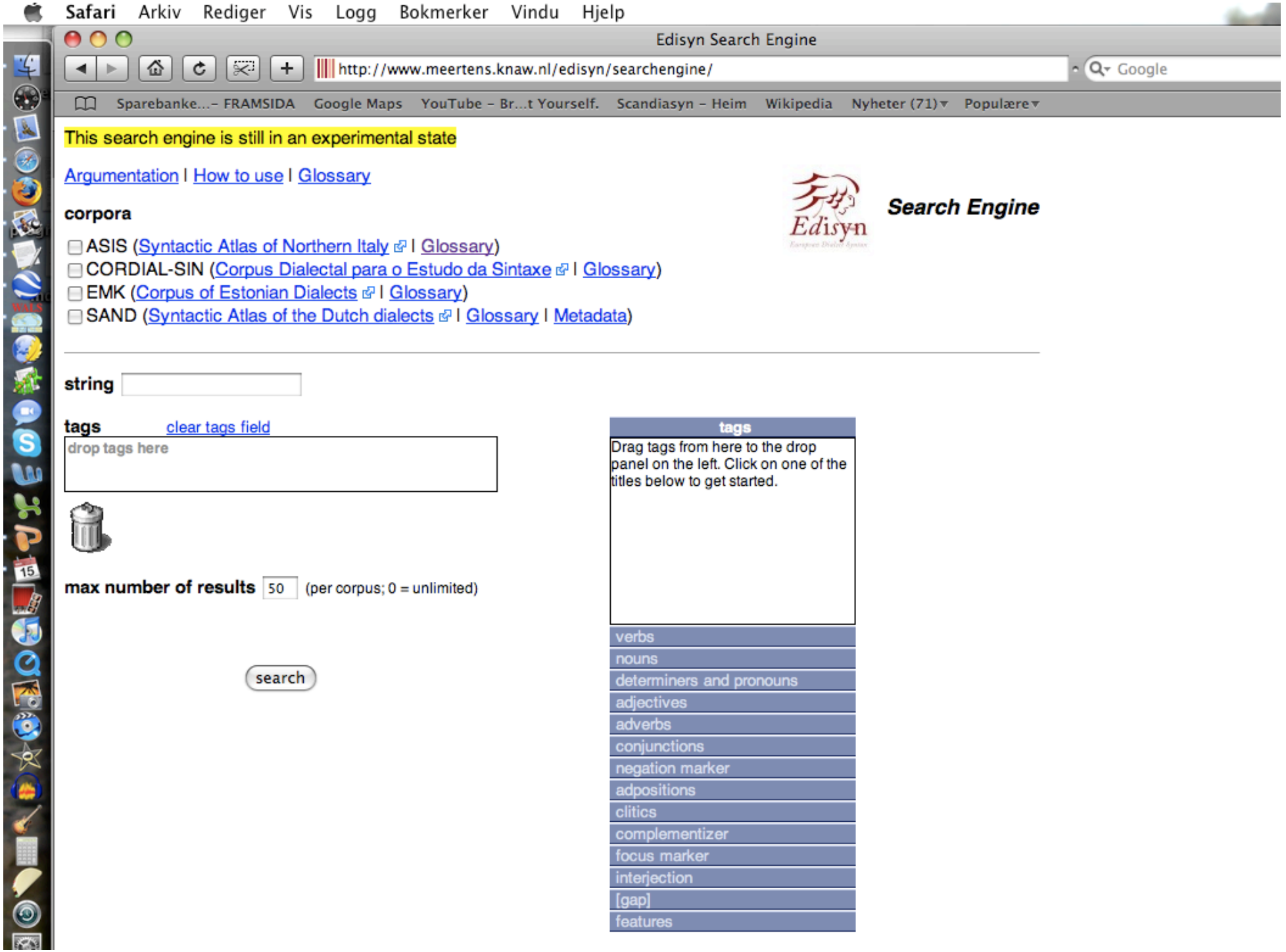
- *The Menota layout and quality hold impressive standards and accuracy; manageable with finite data sets, harder when you want to deal with infinity.*

- *User proficiency in Old Norse/Old Scandinavian required; is automatized glossing a reachable – and desirable – goal?*

  (Google Translate is probably not the savior in this case…)

Edisyn issues

How can you make use of a search engine which combines databases with different tag sets and unglossed examples?


Safari Arkiv Rediger Vis Logg Bokmerker Vindu Hjelp
Edisyn Search Engine
http://www.meertens.knaw.nl/edisyn/searchengine/
Google
Sparebanke...– FRAMSIDA
Google Maps
YouTube – Br...t Yourself.
Scandiasyn – Heim
Wikipedia
Nyheter (71)
Populære
This search engine is still in an experimental state
Argumentation | How to use | Glossary
Edisyn
Search Engine
corpora
ASIS (Syntactic Atlas of Northern Italy | Glossary)
CORDIAL-SIN (Corpus Dialectal para o Estudo da Sintaxe | Glossary)
EMK (Corpus of Estonian Dialects | Glossary)
SAND (Syntactic Atlas of the Dutch dialects | Glossary | Metadata)
string
tags
clear tags field
drop tags here
max number of results 50 (per corpus; 0 = unlimited)
search
tags
Drag tags from here to the drop panel on the left. Click on one of the titles below to get started.
verbs
nouns
determiners and pronouns
adjectives
adverbs
conjunctions
negation marker
adpositions
clitics
complementizer
focus marker
interjection
[gap]
features

Desires for result handling

- *Comprehensible and user-friendly linguistic marking.*

- *Automatized glossing: a crosslinguistic dictionary or the like wanted!*

Cartography – for what?

Maps are legendary in dialectology:

– *display and support diachronic linguistics*

– *display the cultural sides of language*

Typologists love maps:

– *display areal distributions*

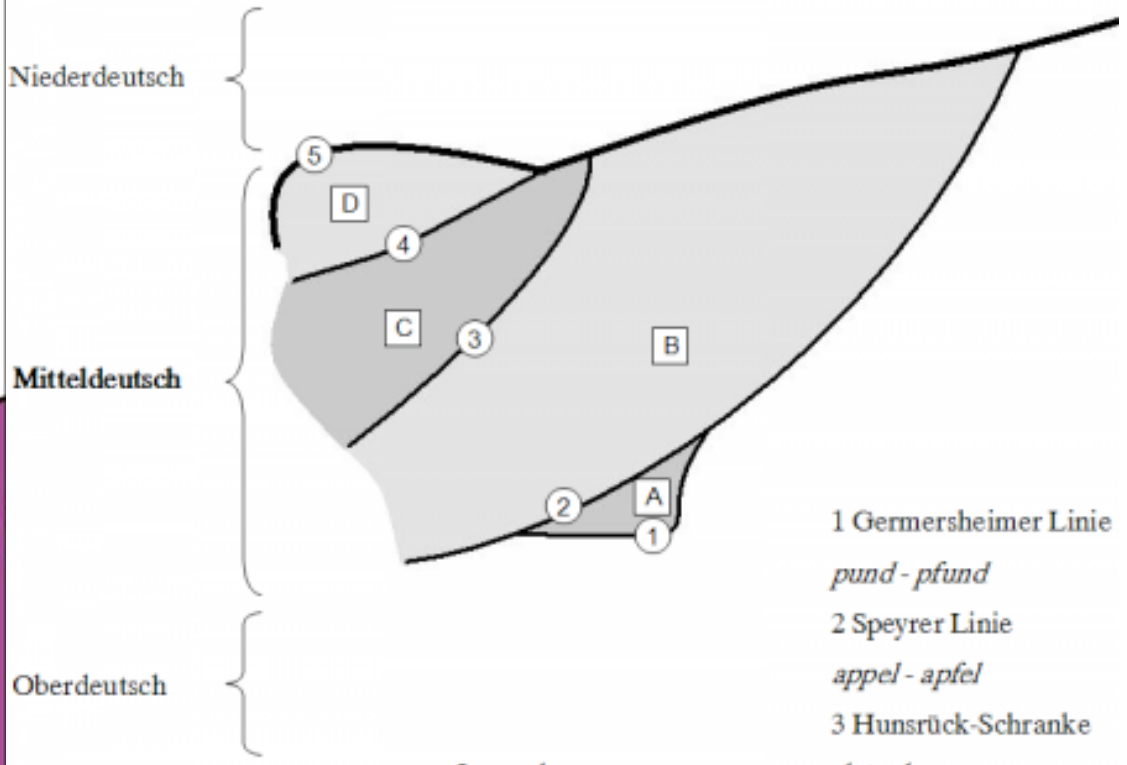
– *(thus) display cultural sides of language*

Maps are useful for theoretical linguistics:

– *display correlations between phenomena*

– *can display details of phenomena*

Der rheinischer Fächer



Legende

A Südrheinfränkisch

B Rheinfränkisch

C Moselfränkisch

D Ripuarisch

1 Germersheimer Linie

pund - pfund

2 Speyrer Linie

appel - apfel

3 Hunsrück-Schranke

dat - das

4 Eifel-Schranke

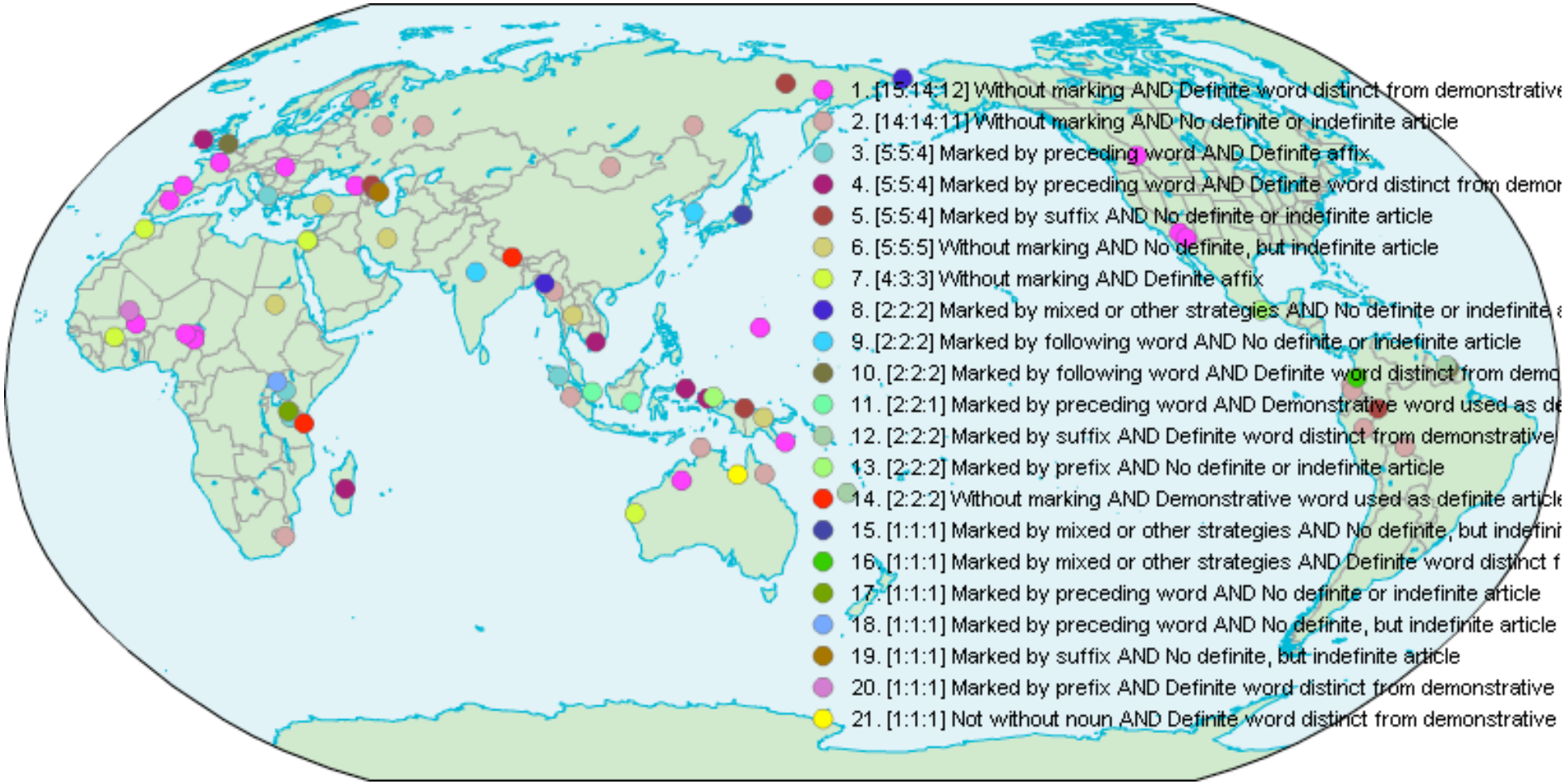
dorp - dorf

5 Benrather Linie

maken - machen

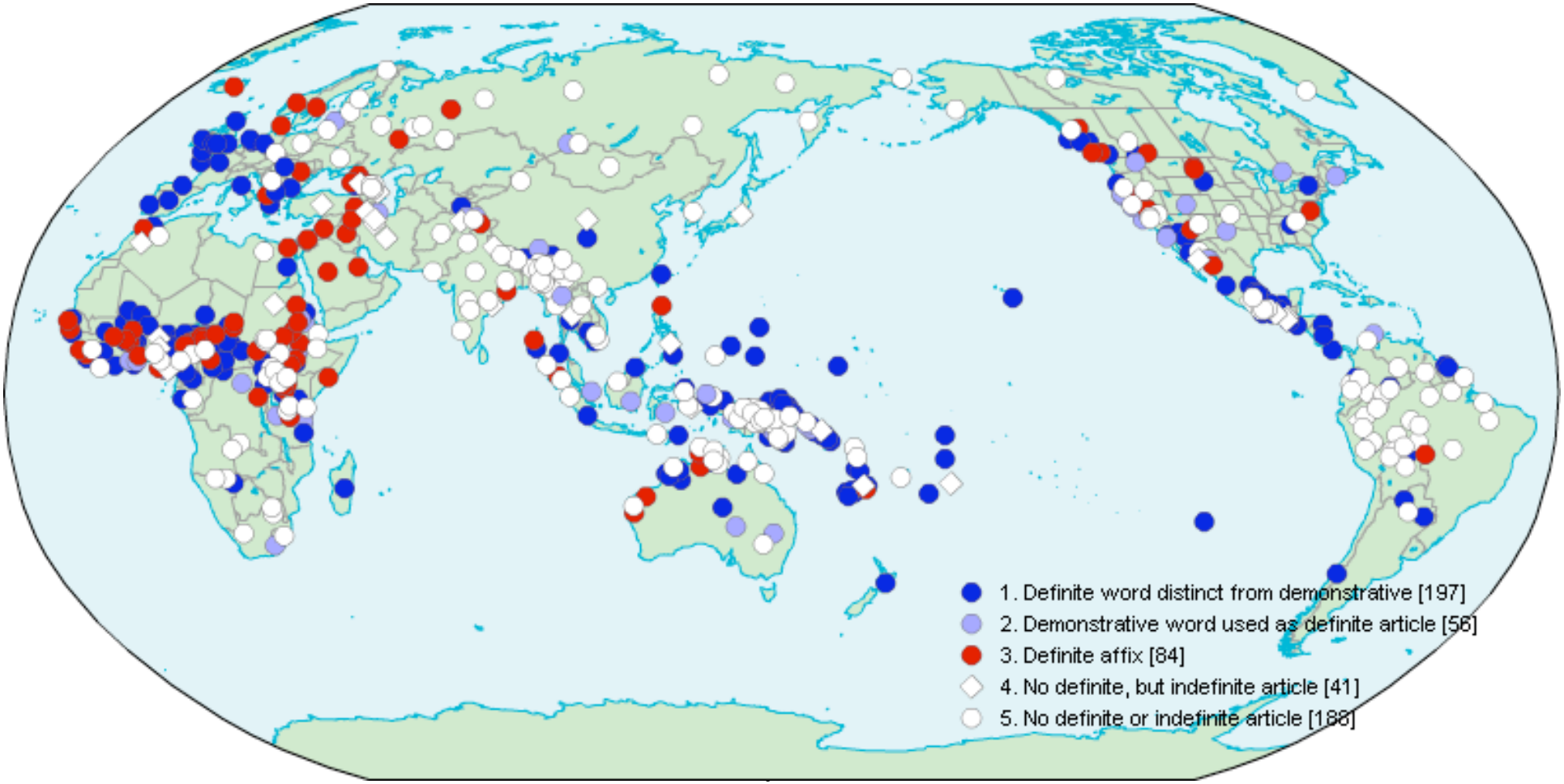
WALS – two feature search

Adjectives without Nouns AND Definite Articles:



WALS – a critical eye

Definite articles



WALS – details to be missed

- The Scandinavian definite affix has different properties than the Balkan (Rumanian, Bulgarian) definite affix.

maður-inn	(Ice.)	*om-ul*	(Rom.)
gamli maðurinn		**batrîn om-ul*	
**gamli-(i)inn maður*		*batrîn-ul om*	

- Mainland Scandinavian also has a "demonstrative word used as a definite article".

den mann-en	**(den) gamle mann-en*	(Norw.)
'that man'	'the old man'	

Seletive global comparison

Hindi-Urdu (Rajesh Bhatt, p.c.):

(1) a. Tum-ne problem **kaise** solve kii (manner)
you-Erg problem how solve do.Pfv.f
'How did you solve the problem?'

b. Tum-ne **kaisii** car khariid-ii (kind)
you-Erg how.f car buy-Pfv.f
'What kind of car did you buy?'

Basque (Ricardo Etxepare, p.c.):

(2) a. **Nola** konpondu-ko duzu kotxe-a? (manner)
how fix-FUT you-have car-DET
'How will you fix the car?'

b. **Nola**-ko kotxe-a duzu? (kind)
how-of car-DET you-have
'What kind of car do you have?'

Adnominal 'how$_{MNR}$' in Norwegian

Tromsø dialect of Norwegian:

(1) **Korsn** ska du løse probleme? (manner)
how will you solve problem-DEF
'How will you fix the problem?'

(2) **Korsn** bil har du? (kind)
how car have you
'What kind of car do you have?'

(3) **Korsn** bil e din? (token)
how car is yours
'Which car is yours?'

ScanDiaSyn questionnaire

Adnominal (manner) *how*:

(1) 'How' car do you have?

→ KIND reading ('bubble')

(2) 'How' car is yours?

→ TOKEN reading ('needle/pin')

Red/redish = bad

Blue/bluish = good

A closer look the south

West:

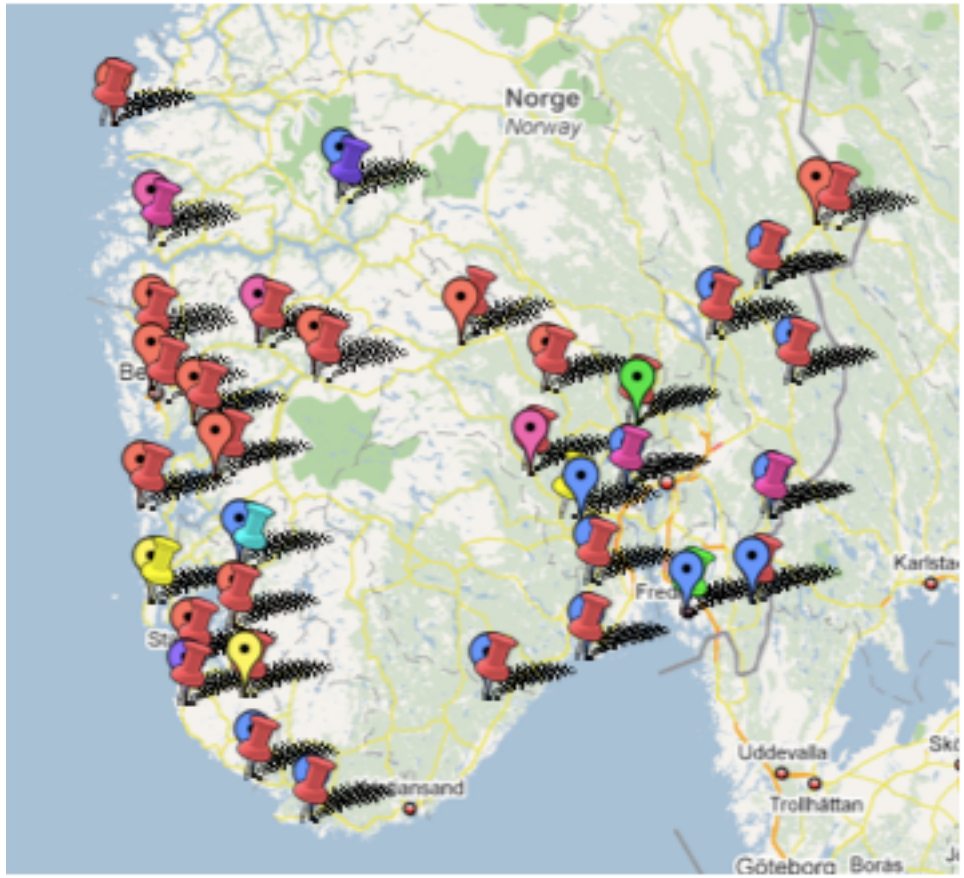
Both readings dismissed to a large extent.

East and south coastal:

Mostly just KIND reading accepted.

North:

Both KIND and TOKEN readings generally accepted.



The adnominal *wh*-cycle

The dialectal information can be paired with the diachronic development of 'which' in Germanic (and 'qualis' in Romance) and 'what kind of' in Norwegian vs. Swedish/Danish:

All show a development whereby query for kind is extended to query for token.

Hence: Maps are useful for visualizing structurally interesting/important facts for linguistic theory.

And vice versa!

Dialect dictionaries and maps

National dictionaries and dialect dictionaries often contain lots of geographic information on each lexical entry.

Huge potential for cartography!

Norsk Ordbok (*korleis* 'how$_{MNR}$')

Rietz' Svenskt dialektlexikon (Projekt Runeberg)

Finally…

- *There are lots of issues to dive into when it comes to infrastructure for language variation research.*
- *Collaboration is essential!*
- *Each project should be true to their own focus: Best practices should arise through collaboration across projects.*
- *The user perspective is important: linguists should see it as an obligation to contribute to infrastructure development.*

Let's get started!