UNIVERSITETET I OSLO
DET HUMANISTISKE FAKULTET

NO-CLARIN fra et UiO-HF-perspektiv

Janne Bondi Johannessen

Nasjonalt møte om CLARIN, Nasjonalbiblioteket, 18.juni 2010

Noen utfordringer om hva som finnes

- Antall ressurser
- Innholdet i ressursene
- Type ressurser

- Infrastruktur for gjenbruk av ressurser
- Infrastruktur for oppgradering av ressurser

Skriftspråkskorpus v/UiO, 2007

Skriftspråkskorpus	Oslokorpuset av taggede, norske tekster, BM	http://www.tekstlab.uio.no/norsk/bokmaal/
Skriftspråkskorpus	Nynorskkorpuset ved NO2014	http://no2014.uio.no/tekster/sok/webconc.html
Skriftspråkskorpus	Oslokorpuset av taggede, norske tekster, NN	http://www.tekstlab.uio.no/norsk/nynorsk/
Skriftspråkskorpus	KAL-korpuset	http://omilia.uio.no/kal/index.html
Skriftspråkskorpus	Usenet-korpuset	http://folk.uio.no/elien/bmsml/veil_LBK.html
Skriftspråkskorpus	Leksikografisk bokmålskorpus	
Skriftspråkskorpus	NP-annotated Norwegian corpus	http://www.forskningsradet.no/servlet/Satellite?cid
Skriftspråkskorpus		http://www.tekstlab.uio.no/grei/
Grammatikkspill/trebank	GREI	
Skriftspråkskorpus	The Oslo Corpus of Bosnian Texts	http://www.tekstlab.uio.no/Bosnian/Corpus.htm
Skriftspråkskorpus	Sidaama-korpuset	http://foni.uio.no/CE2/html/index.php?corpus=omc
Skriftspråkskorpus	Oslo Multilingual Corpus	http://www.hf.uio.no/ilos/OMC/
Skriftspråkskorpus	LOGON Tourist Corpus	http://www.hf.uio.no/tekstlab/prosjekter/tourist/ind
Skriftspråkskorpus	The Sofie Treebank (parallelle tekster)	http://www.hf.uio.no/tekstlab/prosjekter/SOFIE.htr
Skriftspråkskorpus	The OPUS Corpus	http://logos.uio.no/opus/
Skriftspråkskorpus	Samisk korpus	http://omilia.uio.no/glossa/html/index_dev.php?cor

Dvs. 16 skriftspråkskorpus

Nye skriftspråkskorpus siden 2007

- NoWaC-korpuset
  - første versjon av et stort web-basert korpus for bokmål. Denne versjonen inneholder 700 millioner ord. (Emiliano Guevara, Tekstlab)
- RuN-korpuset
  - parallellkorpus med norsk, russisk, engelsk. (Atle Grønn, ILOS + Tekstlab)
- Bibliotheca Polyglotta
  - flerspråklig korpus av historisk viktige tekster. (Jens Braarvig, IKOS + IT-HF)
- SalCorpora
  - Hindi database (Claus Peter Zoller, IKOS + USIT, Tekstlab, U i Heidelberg)
- PROIEL
  - Trebank for gresk, latin, kirkeslavisk, armensk og gotisk. (Dag Haug, IFIKK)
- The French Newspaper Corpus
  - 115 millioner ord fra franske nyhetstekster (Tekstlab).
- Makedonsk tekstkorpus
  - (Tekstlab)

Andre språkressurser: også flere og flere

- Talespråkskorpus
- Taggere
- Databaser
- Ordlister
- Leksikon
- Kartløsninger
- Parsere
- Distribusjonsmodeller for semantikk

Det samlede inventaret av ressurser er IKKE uforanderlig

Endringer i innhold

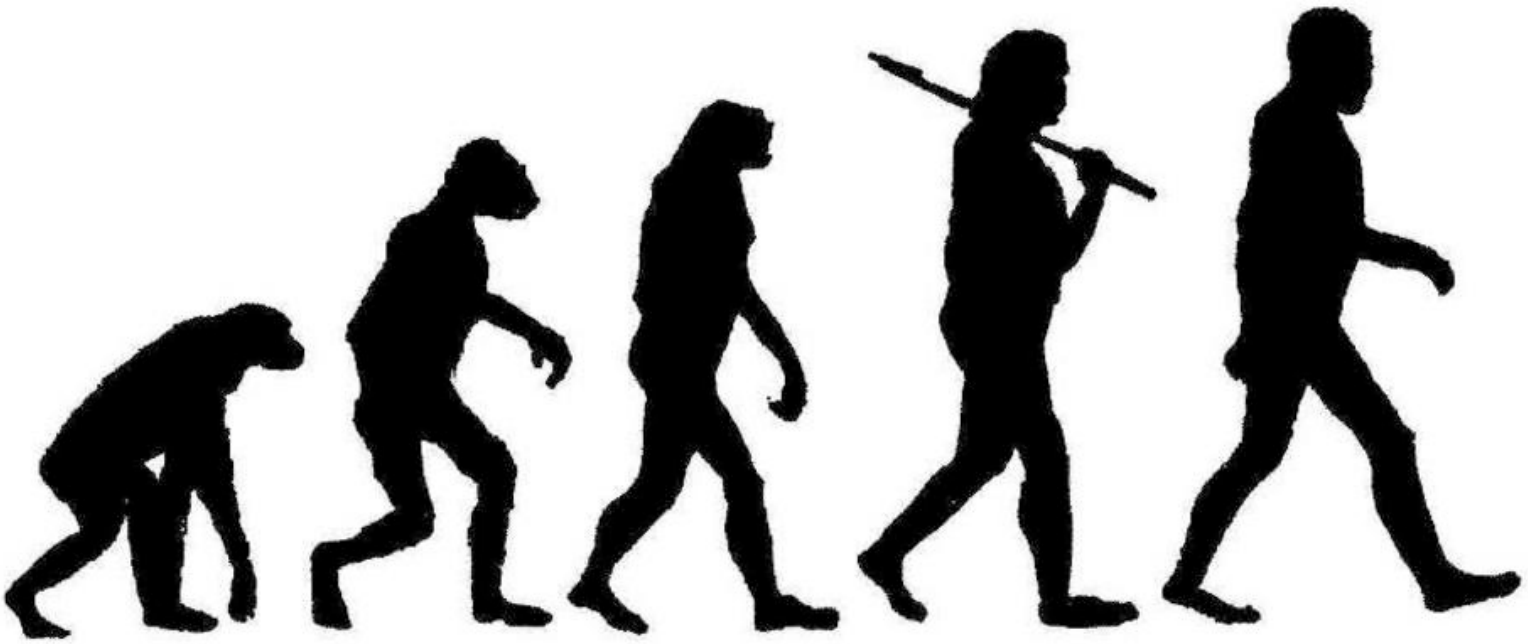
- Nordisk dialektkorpus
  - 2007: noen dialektopptak fra Norge og Sverige
  - 2010: mange dialekter fra Norge, Sverige, Danmark, Island, Færøyene
  - 2012? Også fra svensktalende Finland? Flere opptak fra de andre

  - 2007: svenske dialektopptak var bare med lyd
  - 2010: mange svenske dialektopptak med video også

  - 2008: bare norske dialekter var grammatisk tagget
  - 2010: tagging ferdigstilles for færøysk, islandsk og svensk.

  - 2008: ingen kartløsning
  - 2010: stadig nye kartløsninger i korpuset

Innholdet i et korpus (og andre ressurser) er IKKE uforanderlig



Behovene endres – generelt

- Lagringskapasitet
  - Fra tekst til video: Fra Mbyte til Tbyte
- Akkumulerte datamengder
  - Krever nyere og hurtigere typer analyseverktøy
- Tungregning
  - Fra enkle regelbaserte grammatikker til statistiske beregningsalgoritmer
- Direkte tilgjengelighet
  - Forutsetter web

Stadig nye ressurser på stadig flere steder

- Tidligere:
  - Bare et par-tre institusjoner utviklet språkteknologi
  - Bare et par-tre firmaer var interessert i språkteknologisk råstoff
  - Bare noen få språkfolk og filologer var interessert i bruk av språkteknologi

- I dag:
  - Mange utvikler språkteknologi
  - Mange firmaer er interessert
  - Mange språkfolk og filologer er interessert

Foreløpig konklusjon

- Behov på flere nivåer når det gjelder eksisterende og nye ressurser

Språkteknologisk ressursoppbygging er dyrt – gjenbruk veldig viktig

- Viktig at man vet hva som finnes
  - Behov: dynamisk sentralisert katalog inndelt i temaer, funksjoner o.a. (CLARIN WG 5.6: LRT Integration)
- Viktig at man vet hvor det finnes
  - Behov: dynamisk sentralisert katalog
- Viktig at man vet hvordan det som finnes, er.
  - Behov: detaljerte opplysninger om hver enkelt ressurs
- Viktig at man vet hvordan det som finnes, kan brukes
  - Behov: gode bruksanvisninger
- Praktisk med interoperabilitet
  - Behov: like standarder (CLARIN WG 5.7: Interoperability and Standards, WG 7.2 A: Licensing of Materials)

Infrastruktur: forbedring av eksisterende ressurser (i)

For korpus, leksikon og andre samlinger:

(Clarin WG 5.2: Lexical Resources, WG 5.3: Corpora)

- Web-basert søking
- Nedlastbarhet
- Audiovisuell visning for talespråk
- Transkripsjoner av talespråk
- Grammatisk tagging
- Grammatisk parsing
- Full metadataoppmerking
- Sentral eller lokal tilgang?

Infrastruktur: forbedring av eksisterende ressurser (ii)

- For taggere, parser, analyseverktøy
  - (Clarin: WG 5.1: Tools)

- Bedre presisjon og dekningsgrad
- Web-basert
- Nedlastbar
- Direkte anvendbar

Juridiske problemstillinger

- Rettigheter og plikter ifølge loven
- Er loven for streng?
- Håndheves den for strengt?

CLARIN Virtual Language Observatory - Resources

Demonstrator with IMDI, OLAC, ELRA and CLARIN data (contact: vlw@clarin.eu)

Powered by Flamenco

search

Show tooltip previews of subcategories

ORIGIN

olac (70871)
mpiCorpora (32684)
endangeredLanguages (18478)
cqn (12767)
bas (7419)
lund (5190)
esf (2854)
signLanguage (2768)
dbd (2122)
iLspIntera (1616)
bifo (1521)
ailla (917)
more...

CONTINENT

Europe (54517)
Asia (10647)
South-America (7346)
North-America (6227)
Oceania (2887)
Australia (2817)
Africa (2020)
Middle-America (1268)
Unknown (31)

COUNTRY

Netherlands (20676)
Germany (15604)
Sweden (5701)
Japan (3995)
Belgium (3946)
Turkey (2952)
United States (2872)
Bolivia (2859)
Australia (2836)
France (2794)
Mexico (2733)
Canada (2083)
more...

LANGUAGE

English (26749)
Dutch (19195)
German (14551)
French (4306)
Japanese (4183)
Swedish (4146)
Undeterminated (3650)
Turkish (2768)
Spanish (2605)
Undetermined (1458)
Tzeltal, Tenejapa (1358)
Arabic, Standard (1206)
more...

ORGANISATION

Max Planck Institute for Psycholinguistics (13849)
Bavarian Archive for Speech Signals (BAS) (7419)
Dept. of Linguistics, Lund University, Sweden (3213)
Freie Universität Berlin (1707)
MPI für Bildungsforschung (1515)
University of Cologne (1443)
LABLITA, Dipartimento di Italianistica - Università di Firenze (1442)
German Research Foundation (DFG) (1390)
University of Manchester, School of Languages, Linguistics and Cultures (1349)
University of Leipzig (1333)
Max Planck Institute for Evolutionary Anthropology, Department of Linguistics (1305)
Ruhr-University Bochum (784)
more...

GENRE

Discourse (34370)
spontanous speech (5865)
interview (3213)
Stimuli, act-out (1569)
dialogue (1329)
narrative (1197)
Stimuli (1139)
Movie description (1123)
Singing (865)
Conversation (783)
Elicitation (698)
Unspecified, narrative (523)
more...

SUBJECT

language description (12428)
typology (7502)
generalinguistics (7410)
syntax (7335)
primary text (5480)
monologue about free topic (3909)
lexicon (3905)
phonology (3706)
semantics (3493)
phonetics (2962)
morphology (2614)
people applying for a speechdat prompt sheet via telephone (1956)
more...

T I OSLO
LTET

(f)ekstlab.

This search engine is still in an experimental stage

Argumentation | How to use | Glossary

corpora

- ☐ ASIt (Syntactic Atlas of Italy | Glossary | Metadata)
- ☐ CORDIAL-SIN (Corpus Dialectal para o Estudo da Sintaxe | Glossary | Metadata)
- ☐ EMK (Corpus of Estonian Dialects | Glossary | Metadata)
- ☐ SAND (Syntactic Atlas of the Dutch dialects | Glossary | Metadata)
- ☐ NDC (Nordic Dialect Corpus | Glossary | Metadata)

Edisyn *Search Engine*

---

string

tags clear tags field

drop tags here

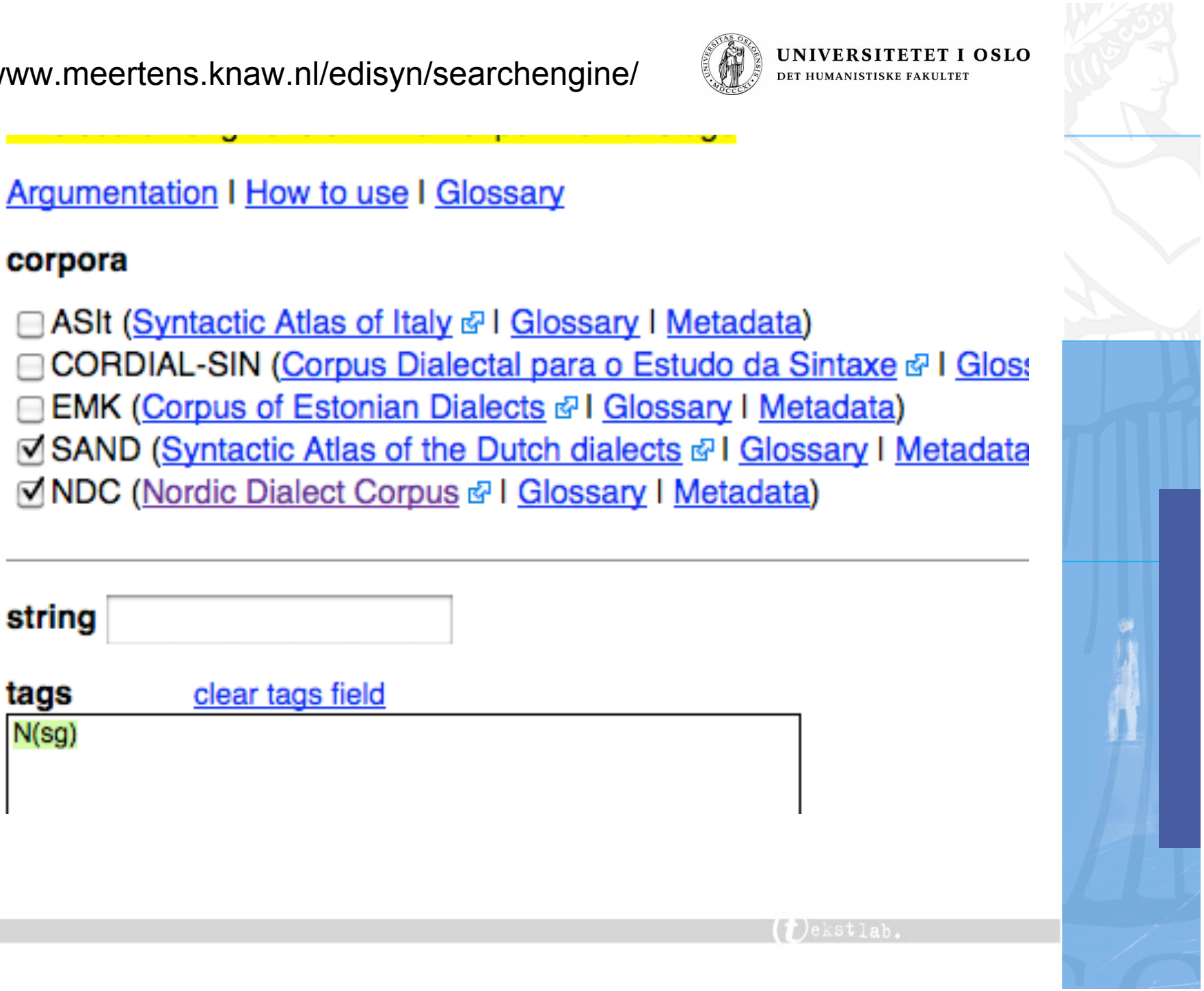
max number of results 50 (per corpus; 0 = unlimited)

search

tags	features
Drag tags from one of the tags lists, or the list of features, to the drop panel on the left (click on the titles below to open the tag lists).	1 2 3 ab abl acc ad add all asp caus com comp coord def dim encl el erg es f fin foc fut gen ger ill imp

verbs
nouns
determiners and pronouns
adjectives
adverbs
conjunctions
negation marker
adpositions
clitics
complementizer
particles
[gap]

http://www.meertens.knaw.nl/edisyn/searchengine/



	zij gelowt dastoe [das toe] eerder in huis$_N$ bins az ik
	she/they believe (2/3 sg.) that you (sg.) earlier in/a(n) house/home are (2 sg.) than/as I
aal_01um	sant ikke sant # em nei for_tida så # går jeg på bygg og anleggsteknikk oppe på e videregående her på Ål
	lemma: bygg, **phon:** bygg, **pos:** subst, **sex:** mask, **num:** ent, **type:** appell, **defn:** ub, **descr:** me, **nlex:** 1

æøå...»

+

−

criteria»

add phrase | delete phrase

Nordic Dialect Cor
ScanDiaSyn/NOR

- Transcription guidelines, transla
- Recording locations
- Transcriptions

Regular expressions: ☐ **Hits per page:** 20 Randomize ☐ Orthographic ◉
Search within: s **Max results :** 2000 Skip tot. freq. ☑ Phonetic ○
Both ○

Search corpus

Reset form

informant +

country + **region** + **area** + **place** +

agegroup + **sex** + **rec (year)** + **genre** +

Show texts

Save subcorpus

Choose subcorpus

Display: **Search within:**

TVÄRSLÅ
Multilingual Dictionary

aal_02uk tenkt å ta det litt spontant (uforståelig)

aal_01um * ikke

aal_01um **sant ikke sant # em nei for_tida så # går jeg på bygg og anleggsteknikk oppe på e videregående her på Ål**

aal_02uk * mm

aal_02uk mm

Trouble viewing video?

context±

Offset

Left -1

Right 1

- Start +

- Stop +

>>

nformants: 524

candiasyn:

WB expression: "([((**word="bygg" %c**))]) ;"

ction : Map

27

esults pages: 1 2

aal_01um sant ikke sant # em nei for_tida så # går jeg på **bygg** og anleggsteknikk oppe på e videregående her på Ål

sannt ikkje sannt # em næi tida så # går e på **bygg** å annlegsteknikk oppe på ee vidregåne hær på ÅL

true not true # em no for the time # so I go on **building** and construction up on e high here in Ål (google)

Spørsmål som var ønsket besvart

- hvordan kan brukere (forskere i humanistiske fag) best jobbe sammen med leverandører (språk- og tekstteknologer)
  - Ut fra erfaring: hyppig kontakt, smidig utvikling og kursing
- hva er en fornuftig arbeidsdeling basert på tilgjengelig ekspertise?
  - …
- hva er mulige oppdelinger av et bredt prosjekt i arbeidspakker?
  - Katalogisering
  - Teknisk infrastruktur
  - Tilgjengeliggjøring av verktøy
  - Tilgjengeliggjøring av korpus
  - Tilgjengeliggjøring av leksika
  - Tilgjengeliggjøringav andre typer data(baser)
  - Juridiske aspekter
- hva er mulige tekniske plattformer?
  - Delvis avhengig av CLARIN-EUs valg, og delvis av hvilke plattformer de eksisterende verktøyene finnes på.
- hva er realistisk å oppnå og hva er store utfordinger som gjenstår?
  - Felles katalog og felles tilgjengeliggjøring er realistisk, felles samkjøring av ressurser fra samme portal er mindre realistisk
- hva kan vi lære av andre prosjekter (som f.eks. CLARIN-NL, D-SPIN osv.)?
  - Bør være noe

tekstlab.

Mange spørsmål!

- På bakgrunn av CLARINs målsetting om tilgjengeliggjøring av eksisterende ressurser:
  - Hva vil det si å tilgjengeliggjøre for innpassing i en overordnet infrastruktur?
  - Hva vil det si at en ressurs eksisterer?
  - Hvor ferdig eller uferdig er en ressurs som kan gå inn i CLARIN(udigitaliserte manuskripter, lydbånd, elektroniske råtekster, rå lydfiler, en halvgod parser, en mangelfull ordliste)?
- Hvor konkret eller abstrakt skal den felles katalogen være?
  - Mens leksikon og korpusdata er ment for teknologer og ikke minst filologer, er det kanskje mest teknologene som kan ha særlig glede av verktøyene?
  - Skal alt være tilgjengelig via en portal?
  - Skal alt være anvendbart via en portal?