UiO : Department of Linguistics and Scandinavian Studies
University of Oslo

WP5: Glossa integration

Anders Nøklestad & Kristin Hagen
The Text Laboratory, UiO

Search engines

The old (i.e., current) Glossa version:

- IMS Corpus Workbench (CWB) the only supported engine
- Tightly integrated with the code base
- Practically impossible to replace with a different search engine such as Korpuskel or Manatee

Search engines

The new version:

- Not tied to any particular search engine – can use any engine for which a Glossa wrapper exists
- Comes with a couple of wrappers out of the box
  - Corpus Workbench
  - CLARIN federated content search
    - Optional module

Search engines

The choice of search engine can be specified for each:

- Glossa installation (e.g. using Korpuskel for all corpora)
- Corpus (e.g. CWB for some corpora and cwb-treebank for others)
- Individual search (for corpora with different types of material, e.g. both CWB-encoded text and syntactic structures)

Search engines

- Creating wrappers for CWB-like engines should be fairly easy
  - Korpuskel, Manatee
- Other kinds of engines will require varying amounts of work based on difference from CWB
  - Tree-based search engines (e.g. cwb-treebank)
  - Search engines for information structure
  - etc.

Glossa My results Log out

« Hide

Category
Collection
Corpus date
ISSN/ISBN
Publication date
Publisher
Publication place
Supercategory
Text id
Title
Translation
Version
Word count

Leksikografisk bokmålskorpus

Simple | Extended | Advanced

han er Search

Glossa My results Log out

« Hide

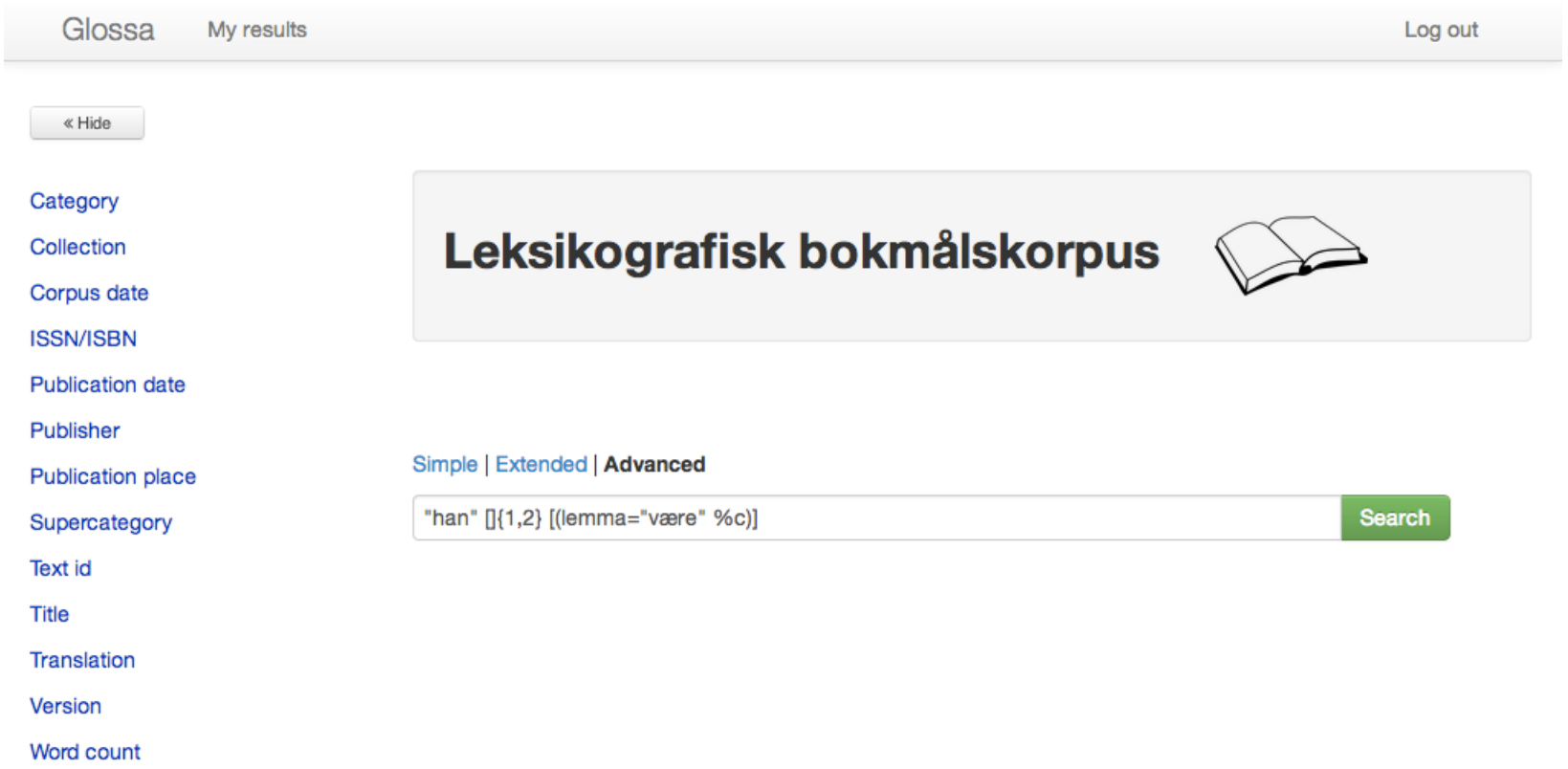
Category
Collection
Corpus date
ISSN/ISBN
Publication date
Publisher
Publication place
Supercategory
Text id
Title
Translation
Version
Word count

Leksikografisk bokmålskorpus

Simple | **Extended** | Advanced

	Interval	
han	1 min 2 max	være
☐ Lemma ☐ Start ☐ End		☑ Lemma ☐ Start ☐ End

Search


Glossa
My results
Log out
« Hide
Category
Collection
Corpus date
ISSN/ISBN
Publication date
Publisher
Publication place
Supercategory
Text id
Title
Translation
Version
Word count
Leksikografisk bokmålskorpus
Simple | Extended | Advanced
"han" []{1,2} [(lemma="være" %c)]
Search

Glossa My results Log out

» Filters New search Found 1999 matches

Simple | Extended | **Advanced**

"han" []{1,2} [(lemma="være" %c)] Search

Save Statistics

«« « Page 2 of 134 pages » »»

AV01Af930343.20	I begge departementer hadde	han tidligere vært	statssekretær , i Utenriksdepartementet fra 1979-81 og i Forsvarsdepartementet fra fra 1976 til 1979 .
AV01Af940001.20	I begge departementer hadde	han tidligere vært	statssekretær , i Utenriksdepartementet fra 1979-81 og i Forsvarsdepartementet fra fra 1976 til 1979 .
AV01Af940006.18	I tillegg kommer det faktum at mye av det	han skrev om var	selvopplevd gjennom flere års fangenskap .
AV01Af940021.43	- Vi er nødt til å se Norges markedsadgang og EUs adgang til fiskeressursene i sammenheng , sier	han . Dette er	en kjent holdning fra forhandlingene om EØS .
AV01Af940039.12	Jeg synes	han her har vært	en dårlig konsulent .
AV01Af940089.43	Men det er påfallende at Olivier Messiaen allerede i 1944 kodifiserte sin musikk , da	han bare var	i midten av 30årene .
AV01Af940130.3	- Hadde jeg ikke kjent faren til Marius litt , hadde	han ikke vært	her , sier Stokke .
AV01Af940164.39	Da det så kom Holen for øre at	han selv var	under etterforskning , fordi det var fremsatt påstander om korrupsjon , ble han så forbannet at han sluttet i politiet

Glossa My results Log out

« Hide New search

Category

Collection

× Aftenposten × Dagbladet

Corpus date

ISSN/ISBN

Publication date

× 1994

Publisher

Publication place

Supercategory

Text id

Title

Translation

Version

Word count

Found 22 matches

Simple | Extended | **Advanced**

"han" []{1,2} [(lemma="være" %c)] Search

Save Statistics

«« « Page 1 of 2 pages » »»

AV01Af940001.20	I begge departementer hadde	han tidligere vært	statssekretær , i Utenriksdepartementet fra 1979-81 og i Forsvarsdepartementet fra fra 1976 til 1979 .
AV01Af940006.18	I tillegg kommer det faktum at mye av det	han skrev om var	selvopplevd gjennom flere års fangenskap .
AV01Af940021.43	- Vi er nødt til å se Norges markedsadgang og EUs adgang til fiskeressursene i sammenheng , sier	han . Dette er	en kjent holdning fra forhandlingene om EØS .
AV01Af940039.12	Jeg synes	han her har vært	en dårlig konsulent .
AV01Af940089.43	Men det er påfallende at Olivier Messiaen allerede i 1944 kodifiserte sin musikk , da	han bare var	i midten av 30årene .
AV01Af940130.3	- Hadde jeg ikke kjent faren til Marius litt , hadde	han ikke vært	her , sier Stokke .
AV01Af940164.39	Da det så kom Holen for øre at	han selv var	under etterforskning , fordi det var fremsatt påstander om korrupsjon ,

CLARIN Federated Content Search

In general each CLARIN center participating within CLARIN-FCS will provide at least the following services:

- provide one or more resources
- support Content-search within those resources
- return search-hits in the agreed-upon format
- support query-expansion (if possible)
- support the selection of a sub-part of the offered resources to perform content-search on that sub-part
- provide support for the sub-part selection by providing CMDI metadata at the same, reasonable, granularity

(from https://trac.clarin.eu/wiki/FCS-specification)

CLARIN Federated Content Search

In general each CLARIN center participating within CLARIN-FCS will provide at least the following services:

- **provide one or more resources**
- **support Content-search within those resources**
- **return search-hits in the agreed-upon format**
- support query-expansion (if possible)
- support the selection of a sub-part of the offered resources to perform content-search on that sub-part
- provide support for the sub-part selection by providing CMDI metadata at the same, reasonable, granularity

(from https://trac.clarin.eu/wiki/FCS-specification)

CLARIN Federated Content Search

- Uses the SRU/CQL standard:
  - SRU: Search/Retrieve via URL
  - CQL: Contextual Query Language

CLARIN Federated Content Search

- Glossa supports FCS in both directions:
  - Locally stored corpora can be searched by external search agents
    - Authentication procedures not yet implemented by CLARIN
  - Glossa can search in corpora stored on other servers in the CLARIN network

Incoming FCS searches

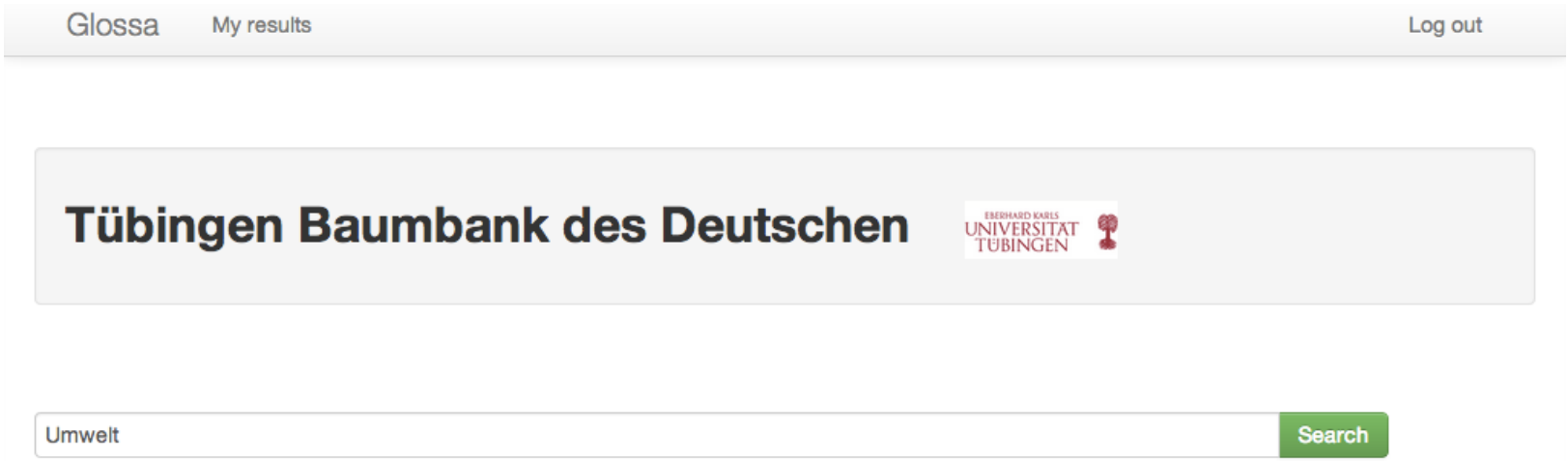
This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<sru:searchRetrieveResponse xmlns:sru="http://www.loc.gov/zing/srw/" xmlns:fcs="http://clarin.eu/fcs/1.0"
 xmlns:kwic="http://clarin.eu/fcs/1.0/kwic">
   <sru:version>1.2</sru:version>
   <sru:numberOfRecords>595362</sru:numberOfRecords>
  ▼<sru:records>
   ▼<sru:record>
      <sru:recordSchema>http://clarin.eu/fcs/1.0</sru:recordSchema>
      <sru:recordPacking>xml</sru:recordPacking>
     ▼<sru:recordData>
      ▼<fcs:Resource pid="http://localhost:3000/fcs/cwb/bokmal#AV01Af930003.4">
       ▼<fcs:DataView type="application/x-clarin-fcs-kwic+xml">
        ▼<kwic:kwic>
           <kwic:c type="left">I 10 centimeter iskaldt overvann møtte</kwic:c>
           <kwic:kw>han</kwic:kw>
          ▼<kwic:c type="right">
             i går 115 andre til nappestrid om mestertittel og pokal ute på Lyseren .
           </kwic:c>
         </kwic:kwic>
       </fcs:DataView>
      </fcs:Resource>
     </sru:recordData>
   </sru:record>
   ▼<sru:record>
      <sru:recordSchema>http://clarin.eu/fcs/1.0</sru:recordSchema>
      <sru:recordPacking>xml</sru:recordPacking>
     ▼<sru:recordData>
      ▼<fcs:Resource pid="http://localhost:3000/fcs/cwb/bokmal#AV01Af930006.15">
       ▼<fcs:DataView type="application/x-clarin-fcs-kwic+xml">
        ▼<kwic:kwic>
           <kwic:c type="left">Bernt Bull sier til Aftenposten at</kwic:c>
           <kwic:kw>han</kwic:kw>
          ▼<kwic:c type="right">
             tar på seg ansvaret for at saken ikke ble behandlet før sommeren .
           </kwic:c>
         </kwic:kwic>
       </fcs:DataView>
      </fcs:Resource>
     </sru:recordData>
   </sru:record>
   ▼<sru:record>
      <sru:recordSchema>http://clarin.eu/fcs/1.0</sru:recordSchema>
      <sru:recordPacking>xml</sru:recordPacking>
     ▼<sru:recordData>
      ▼<fcs:Resource pid="http://localhost:3000/fcs/cwb/bokmal#AV01Af930007.17">
       ▼<fcs:DataView type="application/x-clarin-fcs-kwic+xml">
```

localhost:3000/fcs/cwb/bokmal?version=1.1&operation=searchRetrieve&query=han&maximumRecords=3


```
▼<sru:searchRetrieveResponse xmlns:sru="http://www.loc.gov/zing/srw/" xmlns:fcs="http://clarin.eu/fcs/1.0"
 xmlns:kwic="http://clarin.eu/fcs/1.0/kwic">
   <sru:version>1.2</sru:version>
   <sru:numberOfRecords>595362</sru:numberOfRecords>
  ▼<sru:records>
   ▼<sru:record>
      <sru:recordSchema>http://clarin.eu/fcs/1.0</sru:recordSchema>
      <sru:recordPacking>xml</sru:recordPacking>
    ▼<sru:recordData>
     ▼<fcs:Resource pid="http://localhost:3000/fcs/cwb/bokmal#AV01Af930003.4">
       ▼<fcs:DataView type="application/x-clarin-fcs-kwic+xml">
         ▼<kwic:kwic>
            <kwic:c type="left">I 10 centimeter iskaldt overvann møtte</kwic:c>
            <kwic:kw>han</kwic:kw>
           ▼<kwic:c type="right">
              i går 115 andre til nappestrid om mestertittel og pokal ute på Lyseren .
            </kwic:c>
          </kwic:kwic>
        </fcs:DataView>
      </fcs:Resource>
    </sru:recordData>
   </sru:record>
```

Outgoing FCS searches


Glossa
My results
Log out
Tübingen Baumbank des Deutschen
Eberhard Karls
Universität
Tübingen
Umwelt
Search

Glossa My results Log out

Found 382 matches

Umwelt Search

Save

«« « Page 1 of 26 pages » »»

11858/00-1778-0000-0001-DDAF-D	Auch sein trefflicher Roman " Die Rahl " , der erste in der Folge des zur Lebensaufgabe gestellten großen Romanzyklus , erwählt in der Titelheldin eine große Schauspielerin zur tragenden Gestalt , um die Welt des Theaters einer nicht minder verkleideten	Umwelt	gesellschaftlicher Kräfte im alten Österreich entgegenzustellen .
11858/00-1778-0000-0001-DDAF-D	Denn den Ort , wo alles geschah , will ich melden , weil nur in seiner	Umwelt	das geschehen konnte , was geschah .
11858/00-1778-0000-0001-DDAF-D	Dabei war sie eine sehr kluge Frau , soweit ihr scharfer , kurzsichtiger Verstand in ihrer rings von der guten Gesellschaft umschränkten	Umwelt	reichte .
11858/00-1778-0000-0001-DDAF-D	Und dann tat er dies , der jugendliche Mund , in der ungemäßen Wortmaskerade , die er sich aus seiner papiernen	Umwelt	geliehen hatte , und mit den überschwenglichen Gebärden , die ihm alle Schlichtheit des Gedankens am tiefsten zu verhüllen schien .
11858/00-	Müssen wir also leider auf eine vollständige	Umwelt	und geistige Luft seiner frühesten Jugend bedeuteten .

Installation

- The old version:
  - Installation of required software
  - Installation of Glossa
  - Configuration of Glossa
- For a person with knowledge of the code base: several days
- For others: Very, very hard – practically impossible for most people

Installation

- The new version:
  - Can normally be installed in minutes without any knowledge of the code base
  - A single command installs and configures everything

Installation

Requirements:

- Ruby (version 1.9 or 2.0)
- Some kind of relational database (MySQL, PostgreSQL, Oracle, SQLite etc.)

Installation

- Ruby:
  - Can be installed with a single command
- Database:
  - Many (most?) servers already have a database system installed
  - Normally easy to install using a Linux package manager or Homebrew on Mac OSX
  - SQLite requires no configuration

Installation

- Alternatively use a database running on a different server
  - Very easy to set up: just change a couple of lines in a configuration file

Installation

- Once the prerequisites are met, Glossa itself can be installed with a single command:
  - rails new glossa \
    -m https://raw.github.com/textlab/rglossa/master/app_template.rb
- Adding support for optional modules (FCS, R etc.):
  - One line addition to config file + one command

New use cases

- Institutions can install Glossa on their servers to
  - query their own corpora (via CWB, Korpuskel etc.)
  - let other CLARIN search agents query their corpora (via FCS)
  - query other corpora in the CLARIN network (via FCS)

New use cases

- Even if the institution doesn't have any corpora of their own, they can still use Glossa as a pure client application for searching corpora that are available in CLARIN

New use cases

- Researchers can install Glossa on their own laptops and use it for their own private corpora or as a search client for CLARIN corpora
  - Pro: Can potentially use localStorage or IndexedDB to avoid having to install a database system
  - Con: saved searches will only be available on that particular machine

Other improvements

- Easily stylable
  - Built on Twitter Bootstrap
- JavaScript-centric implementation
  - Yields a snappy, user-friendly experience
- Decoupled front-end and back-end
  - Facilitates functionality such as FCS
- Modular architecture
  - Only install what you need (e.g. FCS and R support are optional modules)

Other improvements

- Statistics via R (http://www.r-project.org/) instead of a set of Perl scripts
- So far only basic frequency lists generated from CQP regexes, but the use of R yields great potential
- Will probably offer the old Perl scripts (or a Ruby translation) as an alternative to provide some statistics options without requiring R

Corpus builder for Glossa

- Generalized building process increasing cross-usability for existing corpora, increasing their overall value
- Workflow implies guidelines for building new corpora
  - Towards CLARIN-conformant source-data (e.g. teiHeader CMDI profile)

Adaptation of preprocessing tools

- Wrappers for taggers in order to ensure consistent behaviour
- Conversions between various internal and external formats used to process corpus data, e.g.
  - TEI, CSV/TSV, JSON, various input/output formats of taggers used at the Text Laboratory)
- Interface sub-module for IMS CWB and CQP query functionality

Current status

- Search in monolingual, written corpora using CWB or FCS is working
  - Database queries that constrain metadata need to be optimized for corpora with large amounts of documents
  - Soon to start beta testing with “real” users
- Easy installation process is working
  - Installation of the *rcqp* R package might be problematic on some older servers

Future work

- Support more of the functionality in the old version:
  - Support for multilingual corpora
  - Sound and video
  - Maps
  - More sophisticated statistics
  - Pruning of search results
  - Annotation (hard to use in the old version)

Future work

- Support corpus versioning (in accordance with CLARIN)
- Possibly make it even easier to install and run Glossa on laptops if it seems like an interesting use case